Psychometric Procedures and Systems Audit:
Data Management Review of Pearson
for the Georgia End-of-Course Test (EOCT)
and Georgia High School Graduation Test (GHSGT)

Richard M. Luecht, PhD

Terry Ackerman, PhD

Center for Assessment and Research Technology (CART)

Greensboro, North Carolina

November, 2009

1.0 Introduction

The CART Psychometric Procedures and Systems Audit (PPSA) is a technical evaluation of the end-to-end *flow* of item and examinee data through an assessment processing system, from test development through scoring and reporting. Each PPSA requires a high level of operational psychometrics expertise covering all phases of examination processing and advanced computer-based testing systems design and implementation expertise. This audit report summarizes the Data Management Review (DMR) phase of the PPSA carried out during on onsite review at the Educational Measurement Group of Pearson offices in Iowa City, Iowa, 23-24 September 2009. The DMR also includes a throughout review of documentation provided to CART by Pearson staff.

The DMR phase of the PPSA is a formal, technical evaluation of documented data systems, data management structures and processing procedures for items, test forms, and examination results—that is, computerized as well as manual systems and procedures used in creating, deploying, administering, and processing an examination. The DMR evaluates the integrity of the data as it flows throughout the Pearson systems, including test-form and item data management, test administration and processing of raw results, back-end examinee results processing psychometric analysis procedures and scoring, software and data management systems improvements, and quality control/assurance steps. The DMR highlights potential weaknesses in those systems and suggests possible solutions or improvements that help ensure the integrity of the data. When and if they do occur, breakdowns in the creation, deployment or processing of an examination usually occur because of minor human errors on otherwise routine tasks or short cuts implemented for efficiency reasons or to resolve a particular problem or exception. When attention wanes on routine tasks or well-intentioned short cuts by-pass established quality control and assurance procedures, errors may result.

This audit report is organized into four sections. The first section outlines the scope of the DMR phase of the PPSA, including an overview of the Georgia DOE assessment programs being managed by Pearson and its subcontractors. The second section contains the evaluation of the test development, data management, psychometric, quality assurance, and information systems and procedures used by Pearson for the Georgia DOE End-of-Course Test (EOCT) and Georgia High School Graduation Test (GHSGT) programs. Issues or potential problems identified by the DMR and recommendations/possible solutions are included in this section. The third section of the report reviews some specific examination problems encountered over the past year and steps taken by Pearson to both resolve the problems and prevent their

recurrence. Finally, the report concludes with holistic evaluation of Pearson's systems and procedures.

2.0 Scope of the Data Management Review

This section of the audit report provides an overview of the Georgia assessment programs being handled by the Educational Measurement Group of Pearson, including a summary of current, relevant contractual obligations. An outline of the primary activities carried out by CART, Pearson, and Georgia DOE staff relative to the DMR phase of the PPSA is also provided.

2.1 Overview of Georgia DOE Examination Programs

The DMR phase of the PPSA covers Pearson data management systems and processing procedures related to the Georgia End-of-Course Tests (EOCT) and the Georgia High School Graduation Test (GHSGT). Pearson's contractual obligations for test development, administration, and processing are also summarized in this section.

The EOCT is administered in grades nine through twelve for eight state-mandated core subjects: (i) Mathematics I (Algebra, Geometry, and Statistics); (ii) Mathematics II (Geometry, Algebra II, and Statistics); (iii) U.S. History; (iv) Economics (including Business and Free Enterprise); (v) Biology; (vi) Physical Science; (vii) Ninth Grade Literature and Composition; and (viii) American Literature and Composition. In addition, legacy Algebra I and Geometry test forms are administered to accommodate students who entered high school under the previously authorized Georgia Quality Core Curriculum (QCC)[1]. Any student taking an EOCT course, regardless of grade level, is required to take the corresponding EOCT upon completion of that course. EOCT scores are averaged in as 15% of each student's final course grade. New EOCT tests are usually constructed for winter and spring administrations. Recycled tests are used for the summer administration and mid-month administrations. The EOCT can be administered via paper-and-pencil assessments or in an online format. Paper-and-pencil assessments are only available during the winter, spring or summer administrations. The EOCT items are developed by Measured Progress (MP) in collaboration with the Georgia DOE staff and high school educators. The test forms are designed and developed by Pearson staff, who also take full contractual responsibility for the

[1] The QCC has been undergoing transition to the Georgia Performance Standards (GPS). The GPS were introduced in spring 2005 in four content areas: ninth-grade literature and composition, American literature and composition, physical science, and biology. The GPS-based social studies EOCT of economics and U.S. history were administered for the first time in winter 2007. QCC is expected to be phased out by the end of the 2010-11 academic year.

following activities: (i) providing comprehensive program management, (ii) overseeing item development by MP, (iii) providing psychometric services including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the EOCT program, (iv) creating customized administration procedures for receipt control, data editing, and scoring processes, (v) designing, printing, and distributing all test materials and ancillary documents, including electronic and Braille test versions, (vi) processing and scanning paper-and-pencil answer documents, (vii) delivering tests and scoring online versions of the EOCT, and (viii) preparing and distributing score reports, both on paper and online within a 5-day turnaround schedule.

The GHSGT is administered for the first time in the eleventh grade and covers five content areas: (i) English Language Arts; (ii) Mathematics; (iii) Science; (iv) Social Studies; and (v) Writing. The Writing assessment is administered each fall; the other four assessments are primarily administered during the spring assessment, with retest opportunities in the summer, fall, and winter. Pearson became the Georgia DOE's contractor for all GHSGT test development activities in January 2007. Prior to 2007, Pearson had been a subcontractor with responsibilities for printing test booklets, student answer documents, and other administration ancillary materials as well as for distributing and collecting test materials. The actual item writing, item content assignments, and answer key verification activities are subcontracted by Pearson to MP.

This DMR audit report does not specifically address data management issues, systems, or procedures handled by MP. A second contractor, the Georgia Center for Assessment (GCA), affiliated with the University of Georgia, is responsible for scanning and scoring all GHSGT answer documents and preparing and distributing score reports. Activities specifically handled by the GCA are likewise excluded from this report.

The contract between Pearson and the Georgia DOE states that Pearson will provide comprehensive program management for the GHSGT, oversee item development with the subcontractor, MP, provide psychometric services, including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the GHSGT program, design, print, and distribute all test materials and ancillary documents, including electronic and Braille test publishing, and prepare and distribute score reports.

2.2 Data Management Review Activities

An audit was strongly suggested by the Georgia DOE Technical Advisory Committee (TAC) in July 2009 in response to several processing error incidents during the previous year. The basic requirements of the PPSA were subsequently outlined during informal discussions between the Georgia Department of Education (DOE) Office of Standards, Instruction, and Assessment and senior research staff at the Center for Assessment and Research Technology (CART). This led to a statement of work (SOW) being prepared, nondisclosure agreements signed, and information sharing arrangements were worked out between CART and Pearson, laying the groundwork for the PPSA to be conducted in three phases between fall 2009 and summer 2010.

The PPSA has three phases: (1) a Data Management Review of documented data structures and formal processing procedures; (2) an Examination Processing Review that focuses on the QC/QA steps actually taken in processing—not just what is documented software programs and manual procedures; and (3) a follow-up Real-Time Processing and Implementation Audit that includes an unscheduled site visit to the vendor's main processing location(s) to observe and evaluate procedures during actual examination processing. This report is limited to the Data Management Review (DMR).

Although the majority of the work involved discussions and various data exchanges between CART and Pearson, the Georgia DOE is the *de facto* recipient of this audit report. In that respect, although Pearson has the right to respond to the content of this report, they do not have contractual or other influence over the conclusions and recommendations made by CART.

The DMR for the Georgia EOCT and GHSGT began with a formal Information Request (IR) provided to Educational Measurement Group of Pearson by CART (see Appendix A). The response to the IR included 198 document files (approximately 200 megabytes) of data provided by Pearson and stored on a secure FTP site to allow rapid downloading of the materials. Two CART staff members were issued credentials and provided access to the FTP site. CART reviewed all of the documentation provided on the secure FTP site.

The DMR also included a visit by two senior members of the CART staff and one representative from the Georgia Department of Education (DOE) Office of Standards, Instruction, and Assessment to the Educational Measurement Group of Pearson facilities in Iowa City, Iowa on 23-24 September 2009. This face-to-face visit proved to be invaluable insofar as CART understanding naming conventions, definitions for, and relationships among system components, procedures and materials provided in

response to the Information Request. The face-to-face meeting also provided the opportunity for serious and open dialog and discussions between CART and Pearson staff relevant to specific aspects of operations at Pearson.

Following the onsite visit to Pearson, CART staff reviewed documentation and conducted their evaluation of the documented procedures. The results of that evaluation, supplemented by notes from the onsite meeting at Pearson, are presented in this audit report.

3.0 Evaluation of Pearson Systems and Procedures

The purpose of the DMR is to identify issues and potential breakdowns in the documented systems and procedures that could directly or indirectly impact the integrity and accuracy of test scores and the reporting of assessment results. The DMR therefore considered all documented aspects of examination processing at Pearson. Pearson staff members were very accommodating in providing information and materials related to the Georgia EOCT and GHSGT, as well as being willing to discuss specifics of their operations.

This section of the DMR audit report summarizes our review and evaluation of eight processing areas: (3.1) test-form and item data management structures and procedures; (3.2) printing, packaging, and distribution of test materials; (3.3) test administration operations; (3.4) post-administration materials handling and data management, including processing of raw results, back-end receipt and scanning/initial processing of examinee results; (3.5) psychometric analyses; (3.6) scoring and score reporting; (3.7) software revision, data management and information systems maintenance; and (3.8) quality control and assurance. A discussion of issues and potential problems is included for each area, as well as recommendations (if any) about solutions and/or possible improvements.

3.1 Test Development: Items and Test Forms

As noted in Section 2.1, Measured Progress (MP) is the subcontractor in charge of item production, editing, and subject-matter content resolution for both the Georgia EOCT and GHSGT examination programs. Contractually, Pearson remains responsible for oversight of the subcontractor and for the final content and processing of the examinations. This evaluation was restricted to Pearson's test development system, nonetheless the recommendations made below may also extend to MP.

As a prelude to discussing Pearson's and MP's item and test development activities, it seems useful to start with some rather obvious distinctions between items and test forms. This distinction is necessary because it highlights an extremely important object-oriented data management principle needed to guarantee the integrity of dynamic database systems. That principle is to always have a single, primary database source for unique data objects[2].

An *item* is a discrete assessment data object with a unique identifier. No two items can share the same identifier, regardless of how similar those items might appear. Every item has a specific rendered form; that is, the particular image of the item that is printed in a test booklet or administered to an examinee on a computerized test form. In addition, each unique item record may contain content coding and other metadata used in test assembly, an answer key (or keys) or rubric information used in scoring, administrative components (special instructions, presentation template links, links to exhibit materials, or computerized presentation or response-capturing components activated at run-time as part of a test resource or definition file), and both classical test theory and item response theory (IRT) statistics used in. Items may be linked, using their identifiers, to a higher-order data objects such as item sets, groups, test sections, or test forms. This object-oriented design perspective allows new assessment objects to be created as needed. For example, an item set merely is a higher-order object comprised of a list of items, an exhibit such as a reading passage or graphic, and a set of presentation rules as to how the item set is presented. It is even possible that the item might be treated as a "super item" and scored as an intact unit using polytomous scoring.

A *test form* is a unique collection of items and other assessment administrative, presentation, and scoring information. Similar to items, each test form should have a singular (unique) identification and representation from a structural perspective. The content of a test form includes, at a minimum, a list of items, a presentation template or formal layout, instructions, timing information, and scoring status information for the items included in the list. Of course, other metadata or auxiliary information may also be included and attached to a particular test-form object. Multiple instances or versions of the same test form should never be active at the same time, unless each version is uniquely identified and functionally treated as a separate test form with appropriate date stamps and version controls.

[2] It is *theoretically* possible to have multiple data sources being synchronously updated—ideally simultaneously—given any change one of the sources.

Pearson obtains the EOCT and GHSGT item-level data from its subcontractor, Measured Progress (MP). Prior to operational use every item goes through extensive content review by Georgia educators as well as editors and test development staff at MP. All items are reviewed for alignment to the QCC or GPS content, perceived difficulty range, clarity of the items, depth of knowledge, correctness of answer choices, and plausibility of distractors. In addition, the content test development specialists review all items for fairness regarding their depiction of minority, gender, and other demographic groups and sensitive topics. These reviews further evaluate passage appropriateness and reading difficulty (for item sets), potential cuing or other interactions between items, reading passages, artwork, graphs, and figures. At that point, Pearson manages the master item database-of-record. All items are further field tested to provide statistical data that is used to evaluate the psychometric quality of the items with respect to multiple criteria (classical test theory statistics such as item difficulty and item-total score correlations indicating sensitivity to the same singular, underlying scale, fit to the IRT Rasch scaling model, differential item functioning, etc.). From a data perspective, these item statistics, item usage, and auxiliary administration information may then be attached to each item record as those data becomes available. Field-tested items go through a second, large-scale data review by Georgia educators to screen the items for the eligible item pool used to assembly future operational test forms. The educators are trained to interpret the basic statistical data when evaluating the quality of the items.

Authorized Pearson staff, MP content staff, and Georgia Department of Education (DOE) staff have [at least review] access for all items in the EOCT and GHSGT item databases via the Item Tracker™ software application. We would expect that any reviews or edits to particular items are directly applied to the data records in the master item database. Once the field-tested items are deemed acceptable for inclusion in the item pool, it is further expected that the items are locked down and any modifications to any data prohibited, except by carefully documented change logs with independent monitoring and sign-off of all changes to any items (or higher-order, intact units such as item sets). We return to this issue in Section 3.1. *It is not entirely clear that item-level changes are made in one and only one place, nor that the lock-downs and quality assurance monitoring of changes are implemented at the item level for the EOCT and GHSGT programs*.

Pearson uses the concept of *test maps* to represent unique EOCT and GHSGT test forms. The Pearson *test map* appears to be the primary data object for most subsequent quality control activities. In this audit report, we use the term *test map* as synonymous

with *test form*. It is helpful to view the status of both items and test maps from creation through test assembly and composition, deployment, test administration, retrieval/return of test materials and response data, psychometric processing, scoring, and reporting.

Test maps are constructed and finalized by joint efforts of Pearson, Measured Progress, and Georgia DOE staff using pools of items that educators have reviewed and considered eligible for operational use. The Pearson Builder™ application is used to select the items from a designated repository for the draft forms. The forms assembly process seems to be well-documented. The item pools for both the EOCT and GHSGT include only field-tested or previously used operational items with item difficulties calibrated using item response theory (IRT) to the appropriate EOCT or GHSGT scale. This allows the test forms to be "pre-equated" to a common scale to provide comparable measurement over time. As part of the test assembly process, items are selected for each test forms to match blue prints that specify the distribution of acceptable content coverage (i.e., proportional representation the test domains and standards) as well as meeting statistical criteria for approximately parallel test difficulty and measurement precision.

Once items are selected for each draft test form—that is, attached to a test map—extensive content reviews are carried out to verify the rendered images of the items, answer keys, and other information included with each test map file. Pearson refers to these as Content Support Services (CSS) reviews. The CSS reviews are initiated by a formal Key Review Request for an entire examination (e.g., for all EOCT forms). As part of the CSS review, the items on each test form are "cold solved" by subject-matter experts, as well as individually checked for content, grammar, correctness of answer keys. Item flags for any apparent flaws are manually input during the CSS review and reported the appropriate Test Map Team (TMT)[3]. The TMT reports the flagged items and associated comments to Measured Progress and the Georgia DOE. Most communication is by telephone or email. Acceptable items require no further action. Measured Progress has 24 hours to respond—per the subcontract statement of work—to all flagged (potentially flawed or miskeyed) items on a particular test form. Georgia DOE subject-matter experts also respond to the CSS review.

[3] One minor improvement recommendation for the CSS review is to add a "source of flag" coded field and flag reason to the key review spreadsheet. The spreadsheet does NOT specify which (if some or all) items are flagged. A second recommendation would be to generate a "summary flag report" that details the types of flags, including flags generated by item analysis results or calibration misfit analyses.

At that point, the test map schedule (printing, etc.) and relevant cost issues are reviewed by the TMT. Per Pearson, and as warranted, scored (operational) items that exhibit flaws requiring repairs to the printed images usually result in stopping the printing and/or halting deployment of the suspect test forms, even if this results in destroying test booklets for a particular test map. Reprinting also typically occurs if flawed items are swapped (i.e., replaced with better items). Obviously, it would be optimal to hold up printing until every flagged is fully resolved; however, practical time constraints sometimes make that impossible to do. Cost considerations obviously enter into the decision making for dealing with very minor flaws (e.g., a missing comma). In fact, if flagging occurs late in the processing cycle, test production may actually go forward with the field test items dropped or modified locally. When that happens, Pearson might put out an errata sheet (e.g., explaining blank pages or other minor changes in a test booklet). In general, late discovery of a flawed or problem item on a particular test map is more serious because fixes are substantially more costly—possibly impacting test administration dates.

The CSS review and TMT follow-up activities result in a final "sign-off" on each test form. Per Pearson, no TMT final reports/authorizations can be issued, in fact, until all item flags are resolved on each test map file. Once the TMT locks down a test map, it is moved to desktop publishing. The test form is composed and layouts prepared for printing.

3.1.1 Potential Issues in Item Database Management

One of the more apparent dilemmas facing Pearson is management and operation of unique, master item databases for the EOCT and GHSGT, especially given the multiple partners in the production and editing of items—that is, Pearson, Measured Progress, Georgia DOE, and any external subject-matter experts involved in the review process. These master item databases should ideally represent each item as a single entity. An item may have multiple images—most commonly if the item is administered via paper-and-pencil and as part of a computerized test—however, multiple images should be properly locked down for each mode of administration in a master item database (e.g., as XML[4] code that cannot be changed). Multiple images within the same mode of administration should not be possible[5]. The same holds for any other data associated with the item (e.g., answer keys, content codes, item statistics, or exhibits).

[4] Extensible mark-up language used for storing digital content in a highly portable, encapsulated format.
[5] High-order assessment objects such as item sets should likewise have a single image at most for each mode of test administration.

Any changes to the item image or item data should occur in the master item database and be propagated to *refresh* all instances of that item or data on various test forms. If this "single source" view of using a master item database is not employed, it is possible to have multiple, different versions of the same item existing within the active testing system[6]. While it is certainly feasible to operate a testing program with multiple instances of particular items, that approach can create potentially serious data management errors as to what constitutes the "official" version. This point is not made to imply that Pearson allows multiple instances of the items as standard operating practice. That is certainly not the case. Rather, we wish to emphasize the need to carefully scrutinize any procedures that by-pass a "single source" system by allowing changes to anything other than the master item database records.

The master item database for the EOCT and GHSGT items appears to be accessible via Pearson's Item Tracker™ software application[7]. Item Tracker can query and summarize the test content and item counts by standards-based codes, item versions, rationales, rubrics, scoring documents, source documentation, metadata, and, statistics. The item data in Item Tracker also includes all of the rendering information including stimuli (e.g., passages for problem-linked item sets), item stems, and distractors. Item Tracker provides subject matter experts and other authorized users with online access to the items. As noted above, the software also provides basic inventory reporting functions for tracking item counts by relevant fields. Pearson puts the items received from MP into the database underlying Item Tracker. In turn, Pearson's Builder software is used for test forms authoring and to subsequently generate the item lists for the test maps.

It is not clear, however, that changes made in Item Tracker are formally *locked down*—prohibiting changes of the primary item record—and then propagated into every test form (map file) to refresh the images, answer keys if relevant, etc.. There is an Item Bank Change Log generated in Item Tracker, which is good. But that Change Log only shows gross counts of changes and it is not apparent that the counts are reconciled to authorized and finalized changes to the item database. The final quality assurance for items instead appears to be principally handled via the test map proofing and verification process (i.e., the generation of the test definition specifications that effectively act as form-level resource files). It appears that this quality assurance

[6] A common misconception in database systems design is that merely creating unique identification numbers for each record or data object guarantees the integrity of the data. If the same, unique identification number, however, points to two (or more) different sources of the data object, the integrity of the data is obviously compromised.

[7] We say "appears to be accessible" because it is not clear whether Item Tracker is directly accessing the master item database or, rather, an interim extract from the database.

process starts at the test construction process, not in Item Tracker *per se*. That is, although items can be reviewed in Item Tracker and modified, the majority of review activities appear to occur after draft test forms have been created. To the extent that the test map "detaches" an item from the master item database, allowing subsequent changes to be made in desktop publication or by otherwise manipulating the printing of the test booklets, there may be a violation of the aforementioned principle regarding single-source data structures. If two rendered versions of an item exist under the same identifier—one in the database underlying Item Tracker and the other implemented in the image file attached to the test map—there is some potential for serious errors to result.

In theory, the test map data-object model inherently controls the versioning of items as they appear on a particular test form (map). However, it seems possible that a single item could exist in multiple versions (i.e., given changes in image, answer keys, statistics, content coding, or other data associated with the test map file). The designated master item database as a single-resource would resolve this issue, even if only used to routinely reconcile and report discrepancies between the master image, answer keys, etc., with instances attached to the proofed, final test map record(s). These types of reconciliation/discrepancy reports might enhance the versioning control over the items. The Change Log in Item Tracker needs to be 100 percent reconciled with changes to the master item database. In fact, it seems reasonable to suggest that all discrepancies between any instance of an item on any active test form and the master item database should require resolution with dual sign-off and verified lock-down of the final changes in the master item database.

It is interesting to note that item statistics (field-test or operational) do not directly appear to automatically generate flags, except as reported outputs from the item analysis (IA). That is, the IA flags *per se* do not appear to be stored in the master item database, along with resolutions. Rules for flagging items are clearly dictated by the contractual requirements between Pearson and the Georgia DOE, but best practice would suggest that, regardless of which flags get generated, they should be uploaded to the master item database—at least following a "final" IA. Examples would be flags for extreme difficulties, low or negative item-total test score correlations—possibly signaling miskeyed items, IRT model misfit, or incorrect distractors with positive response associations with total test scores (again, possibly signaling miskeyed items). That is, it may be useful to store ALL flags in Item Tracker to signal a content review whenever those items are chosen on future test forms.

We also recommend reporting in Item Tracker counts of various types of flags and resolved flags as a standard query-based report. For example, a flagged-item status

report might show counts for the entire item pool or for a particular test map, indicating: (i) flagged items; (ii) resolved flags; and (iii) pending flags. The number of resolved flags should obviously equal flagged item count and pending flags should be zero before. The statistical flags could be auto-loaded to Item Tracker—that is, automatically updated in the item database, but individual flag resolution cannot be automated (i.e., every flagged items MUST be looked at). It would appear that summary reports for all types of item flags, from content-based reviews as well as statistical flags, could be readily stored in the primary database. Similar reports might be added to test map tracking to prevent flagged items from getting too far into the production cycle.

3.1.1 Potential Issues in Test Form Management

Similar to items, each test form (map file) should have a singular identification and representation in a primary database. In that regard, test forms should never be viewed as amorphous entities that undergo continual changes. The content of a test form includes, at a minimum, a list of items and item sets, and associated assessment data objects (e.g., exhibits such as formula sheets), a presentation template for layout, instructions, timing information, scoring status information for the items included in the list, and an approved image of the test form (e.g., digital published version, blue-dot or blue-line camera-ready copy of booklets, etc.). Once a test form is finalized and moved into production, its status changes to "locked down." No changes to any items, instructions, passages, exhibits, or administrative controls (e.g., timing conditions) are allowed[8]. There are two options for dealing with changes to test forms after the production sequence begins. One approach is to back the affected test form(s) out of production, make the changes, and then restart the production process from the beginning. The other approach is to create a new version of each changed test form. Certainly, once a particular test form is administered, if it is later revised, the second version of the test form becomes a new test form.

As previously noted, Pearson uses the term "test map" to refer to a particular test form. The test map concept embodies what others have called "test definition" or "test resource" files. The important distinction is that there is a one-to-one correspondence

[8] Certain testing programs use the concept of "slots" that can be randomly or systematically filled with pretest items, external-link equating items, or other content. Each unique collection of items is actually a different test form, regardless of whether they are formally treated that way throughout the processing cycle. Failure to recognize the unique relationship between a single test form and the items that comprise that form—scored or not—can result in serious data management breaches.

between a test form and a test map file[9]. In that regard, each test map instance appears to be uniquely identified and is created as part of the test construction process. The test map records are monitored by the Test Map Team (TMT), with many quality control steps involving content reviews and other checks on the printed booklets.

One of the more challenging for managing test maps appears to involve changes to the items on the form that may become necessary late in the production cycle—possibly after printing has started or even finished. Ideally, any changes to a particular item should occur in the master item database, be locked down there (including the image, if altered) with appropriate sign-offs and change reporting, and then propagated to all forms (test map files) that use that item. Affected forms that are already in production would be backed out of production. The necessary changes would be made and the production sequence restarted. There are recognized, pragmatic costs that make this "optimal" approach as somewhat unappealing from a financial standpoint (e.g., the cost of reprinting thousands of test booklets because of a minor typo on a field-test item). Nonetheless, the principle of "single source" for the items cannot be sacrificed by "detaching" the items from the master database, once they become part of a test map. At the very least, a mandated reconciliation and comparative analysis should be routinely performed to determine and report any and all discrepancies between the final test map file and the item images and any other item (or item set) data stored in the item master data base. This type of report would provide an additional quality assurance mechanism to ensure that late-cycle changes to the items are back-propagated to the master item database. A follow-up reconciliation and comparative analysis could verify changes were synchronized in both data sources (the master item database and the test map resource file).

Another potential area for problems relative to test map integrity involves data that is associated with the test form, not particular items. Examples would be title and instruction pages, formula sheets, or other reference materials either printed in the test booklets or otherwise administered as part of the testing package. The potential for errors occurs if there are multiple test forms and one of more of these test-form-level data objects is changed late in the cycle. For example, what happens when a formula sheet is used on two test forms, but only changed, even in a minor way, for one test map file? For multiple test forms, we can conceptualize a test map in an object-oriented

[9] Some computer-based testing companies refer the "test resource" or "test definition" file as a collection of all active test forms—whether preconstructed or adaptively constructed in real time. A test map is understood here to imply the unique set of all items and/or item sets administered to one or more examinees.

manner, we can have a master test map instances that fixes certain elements for every test map instance (e.g., fixing the inclusion of instructions, formula sheets, or other exhibits). Any changes to those assessment data objects are locked down at the master instance level. In turn, each test map instance becomes a unique test form that inherits the fixed elements from the master, precluding those essential fixed elements from being inadvertently changed for a specific instance of the test map. An alternative to this approach of using a formal higher-order test map object for multiple forms would be to automate comparisons across multiple test forms. It is not clear whether this type of automated checking with required actions is provided via the Pearson *Key Review Report*[10], which checks for multiple instances of answer keys, or as part of the *Repeated Items Report,* which lists the items which appear on more than one form (including additional relevant item and passage information). The latter two reports appear to be somewhat passive reports that require only cursory examination. A better approach would be to require formal sign-off on all item-level discrepancies across all test maps, as well as verifying the final versions of individually changed items are appropriately updated in the master item database and then locked down.

3.2 Composition, Publishing, Printing, Packaging, and Distribution of Test Materials

The composition, publishing, printing, packaging, and distribution of test materials includes extensive review of every published test forms with explicit version controls (see, for example, the Pearson documents: *Printing Blue Dot Check-off List* and *CR Packaging Quality Check Standards*), bar-coding and other automated or semi-automated quality control procedures for materials handling, and extensive training of the test administrators and end users.

Publication essentially involves a hand-off of the "locked" test map file. Each test map file for paper-and-pencil tests is processed by a Pearson application called BookMap™ that pulls the items for composing and desktop publishing each test form booklet. XML coding of the content and Apple scripts are used to generate formatted PDF files for each item, using template and/or style sheets to generate the layout files used by BookMap. BookMap only allows one unique item number (UIN) per item to at least prohibit duplicate items from appearing in one booklet. Item sequencing is run as a script for each test map file. The source of these item pulls is not specified in the Pearson documentation, but is assumed to be the master item database, rather than an interim item database. Per Pearson, each test map file must be locked down with

[10] The simple fact that this *Key Review Report* exists suggests that it is indeed possible to have multiple keys for a single item within different test maps or other intermediate files used in processing or analysis.

double sign-offs from the test map team. Minor booklet exceptions are allowed if they are discovered late in the production cycle—for example, passage-linked items with change of a comma to one item in the set, or running copies of common page and inserting it on multiple forms. Test form booklets are then moved to publication and loaded into Tracker.

Computer-based (online) tests follow a similar set of composition and publishing steps. A series of software tools are available for the computer-based test forms (referred to as "eTesting" by Pearson), including Form Review Authoring™ application, Test Nav™, and Form Tracker™. These tools mandate a series of authoring steps and quality control reviews, prior to release of any test form to the field. First, the UINs are locked down for each test map, analogous to the book map and loaded into the Form Tracker application that monitors the activities of all groups that see the form. The items go through additional editing steps for online rendering and presentation (e.g., adding hot spots around the response control, removing boxes not needed for online presentation, etc.). Note that this editing is apparently not locked down in a master database as a fixed image (or mark-up) of each item. Rather, each computer-based form is manually composed and edited by a designer who uses Georgia EOCT or GHSGT specifications for formatting, checks, and signs off. A project planner (member of the Test Map Team) repeats the checks and also signs off. The computer-based test (CBT) forms are initially published to in a quality control environment and again manually checked for functionality in a WYSIWYG[11] version using the Pearson application called Test Nav. There is a 100 percent proofing of the computer-based test forms to the hard-copy test booklet insofar as the look-and-feel correspondence between the items. The CBT items and attributes are then loaded to the Form Tracker application where the form may undergo additional, iterative proofing by the TMT as well being made available to Georgia DOE staff for additional proofing and sign-off.

Quality control forms include be sign-off for every published paper-and-pencil test booklet and every CBT form in the Pearson form authoring system. Once a test form is signed off, it is locked down for publication and subsequently moved into production—that is, printing and shipping for paper-and-pencil tests or, published within the online test administration environment for computerized test delivery.

Once a test map is moved to printing, the majority of the processing at Pearson appears to be automated or semi-automated. Each printed test booklet is sealed using paper tape and a tracking the bar code ink-jet printed onto the forms. Booklets are then

[11] *What You See Is What You Get*, referring the to onscreen look-and-feel of each item.

packaged and sealed for shipping. Scanning technology is used to verify all materials handling. Pallets of test booklets and answer sheets are verified from the sales order (or back-ordered). The scanners lock if something is missed.

Test booklets are shipped using Pearson-generated shipping labels for Federal Express with 100 percent quality control verification done on all shipments. Mis-shipments, lost shipments, and other anomalies are recorded and stored in an Oracle database. Pearson claims that their current best practices are better than minimum contractual requirements. No counter-evidence to that claim was evident during our DMR visit to Iowa City.

3.2.1 Recommendations for Publishing

Final item edits are presently handled in the publishing phase of test form production and necessarily incorporate different software authoring tools for the paper-and-pencil versus computer-based test forms. Although 100 percent QC checks are conducted to verify each item and to compare the paper-and-pencil booklets to their online counterparts, there are only *cursory* checks across multiple forms, should the same items be used on two test maps[12]. This issue could be resolved if, as noted in Section 3.1.1, the master item database contained one or more finalized images of each item (PDF image files or XML coding, associated references to Apple scripts, and other relevant presentation data). As an alternative, a comprehensive, comparative reconciliation of the master item images (and other data) against the final published images in the publication and final production could highlight discrepancies and require synchronization steps to ensure that the master item database is updated with the "official" [best] image for each mode of administration. Dual sign-offs should be required for all changes to the master item database and/or test map files.

A second recommendation is to incorporate both the [requisite] paper-and-pencil booklet desktop template references (script IDs, style sheet IDs, etc.) and the eTesting authoring and scripting references into the test map specification. It should be possible to actually restrict the publication editors from accessing style sheets or template objects not specified in the test map master. (Also see recommendations in Section 3.1.2.)

A final recommendation is to record all editing and publication exceptions as part of the test map object file, including records of any sign-offs. Although this type of

[12] Common items across test maps are reported in the *Repeated Items Report* and multiple versions of the answer keys on different test maps are reported in the *Key Review Report*. The timing of these two reports within the production cycle and who sees/acts on them is not apparent from the Pearson documentation. Nor is it clear that the final changes are made to the master item database, locked down, and signed-off.

exception data may currently be stored in Tracker, as noted earlier (see Section 3.1.2), each test form (map) should have a single source repository for all information related to every test form.

3.2.2 Recommendations for Distribution

As might be expected, the end-of-cycle printing, packaging, and distribution processes used by Pearson draw heavily on the company's extensive experiences in test booklet printing, packaging, and shipping. Their materials handling practices also appear to adhere to standard best practices used in manufacturing and many other industries.

Although reconciliation of the test materials shipped versus what is received and ultimately returned by the Georgia district test coordinators has not been a problem, we do recommend that Pearson broadcast to the districts the shipment box count and the tracking number for the shipments. The district test coordinators could then enter the receipt information online to verify the shipment, prior to the test administration (i.e., implementing a type of required end-user verification that might be more detailed than a shipment delivery confirmation from Federal Express).

3.3 Test Administration Operations

Most all of the EOCT and GHSGT administration activities are handled by Georgia test coordinators and other school district personnel. The district test coordinators are considered to be liaisons to the Georgia DOE Assessment and Accountability Office and are responsible for distributing test administration information and test materials to each school within each district. The test coordinators work closely with the EOCT and GHSGT DOE program managers. The Georgia DOE conducts mandatory workshops for all test coordinators, including a System Testing Coordinators Conference, a Pre-Administration Workshop, and specific training regarding local scanning, examination record changes, and online testing (where warranted).

3.4 Post-Administration Materials Handling and Data Management

Post-administration materials handling begins with the return of test materials, physically or electronically, to the processing facilities, subsequent scanning activities (for paper-and-pencil answer sheets), and uploading of the scored response data. These processes only apply to the Georgia EOCT. The post administration management and

scanning of answer sheets[13] for the GHSGT is handled entirely by the Georgia Center for Assessment (GCA), a research center affiliated with the University of Georgia. Our review does not address activities carried out by the GCA.

All paper-and-pencil return materials are received at Pearson and identified by special labels that had previously been sent to the district test coordinators. Each receipt package is scanned in at the dock. The packages are then opened and receipt information recorded, with a reconciliation sent to program team that includes counts of completed booklets and answer sheets, incompletes, unused test materials. Incomplete shipments are handled as exceptions.

The data preparation staff at Pearson roughly compare (i.e., visually) the teacher and/or test coordinator prepared header inventory sheets to the actual counts of materials received. Per Pearson, the headers sheets are usually prepared quite well for the Georgia examinations and there are not many damaged documents.

Received materials are then moved to scanning and prepared in batches. Prior to processing an examination, the Test Map Team (TMT) provides the Pearson scanning group with test decks for every paper-and-pencil test form. These test decks are used for scanning engineering program verification. Results from test deck are compared to expected results and the scanner set-ups are modified as needed.

Batch creation starts by logging the answer sheets in the work-flow management system. Answer sheet batches are probed for moisture content to avoid sheets sticking together. If the moisture level is unacceptable, the batches are moved to acclimation and allowed to dry out at least 8 hours. The answer-sheet batches are slit and then "flip-and-jogged" to align the pages. The batches are then read into the scanner which scans and checks the header coding and the sheet counts. Unique identifiers are assigned for tracking.

The scanned data is then moved to the Pearson imaging capture environment (ICE). ICE counts timing and alignment tracks and flags out-of-tolerance batches. Flagged batches are sent to editing. Gridding errors, etc., are manually keyed in (e.g., missing middle initials). When header count does not match the scanned and processed count of answer sheets, the program team may be informed. The final, end-count checks are all fully automated and securely processed with a 100 percent reconciliation

[13] Pearson provides GCA with an answer key layout file that lists the subject area, domain, booklet position, and status. Interestingly, this key file does apparently (per Pearson documentation) does not include the item ID. In that regard, we are left to trust that the item positions are a fool proof way to match answer keys to items for the GHSGT.

of the examinee response data records scanned/edited to the header sheets. Scanning accuracy checks appear to use proportional, systematic sampling as an additional quality control step. Audited records present the scanned answer document image and the data record for manual review. The editing system messaging is also monitored for error conditions that signal potential problems with one or more scan sheets.

3.4.1 Recommendations for Materials Handing and Data Management

The Pearson materials-handling system appears to be fairly automated and uses standard best industry practices for receiving, processing, and reconciling the data. Perhaps the only loophole is that, from a data perspective, a particular examinee record is first created when an answer sheet is received at Pearson and scanned into the system. Up to that point in time, all reconciliations are based on count checks, match to headers on the return packaging. Those count checks seem adequate to grossly account for all materials shipped to a district and returned (completes, incompletes, and unused test materials).

One suggestion for possible improvement—albeit, probably a non-feasible one given the current systems capabilities and budgets for the EOCT and GHSGT programs—would be to create a primary source for examinees and reconcile the receipt or absence of examination records (both scanned and computer-based records) to that primary examinee record source. This type of reconciliation system is sometimes used in computer-based certification and licensure testing settings where "eligibility files" containing all viable examinee identifiers provide the primary source data for reconciling to scheduled and completed examination records, including exact identification or no-shows, corrupted records, duplications, or even lost records. Obviously, this type of primary source documentation would need to be prepared by the districts, verified, etc.. The *Classroom ID Sheet* noted in the Pearson packaging materials list might serve this purpose, but would require electronic entry against which to match the subsequent answer sheets. In making this recommendation, we acknowledge that the risk of error seems low since Pearson has not experienced any recent problems with lost records due to scanning or processing of the raw test materials.

3.5 Psychometric Analyses

The Pearson psychometrics group is involved in the following activities: (a) test construction to ensure that test forms meet statistical targets; (b) preparation of scoring tables for each test form using "pre-equated" item response theory (IRT) statistics; (c) data management for psychometric analyses; (d) item and classical test analyses,

including distractor analyses; and (e) item response theory calibrations and equating for operational (scored) and field-test items. These issues are correspondingly covered in Sections 3.5.1 to 3.5.5. Our recommendations are then presented in Section 3.5.6.

3.5.1 Test Construction

Formal test construction specifications for the Georgia EOCT and GHSGT are provided in documents forming part of the functional requirements outlined in the contract between the Georgia DOE and Pearson. Aspects of test construction and composition were also covered in Sections 3.1 and 3.2. The Pearson psychometrics group has two primary involvements in test construction. The first involvement is to assist in the screening of items included in the eligible item pool. Ideally, subject to the available inventory of items, only items that are in the proportion-correct range of difficulty from .3 to .95, have moderate discrimination or better (r_{pbis} > .25), and no differential item functioning flags for relevant population subgroups are included in the eligible item pool.

The second area of involvement is to ensure that the approximate statistical targets for each test form are met so that the observed score distribution stays relatively stable over time (assuming no dramatic changes in curriculum or performance by Georgia students). The item response theory (IRT) framework used for calibrating and equating appropriately adjusts the students' scores if they take a slightly easier or more difficult form than other examinees. However, it is still best practice to construct the tests to provide essentially equivalent observed scores, especially since raw score cut points and related information are routinely reported to educators, parents, and other constituents in Georgia. The near statistical parallelism of the observed scores can be [approximately] achieved by matching target mean and standard deviation values of item *p*-values and IRT Rasch item difficulty estimates, as well as average point biserial correlation targets. Pearson further attempts to match target IRT test characteristic curve values (especially at the cut points). The actual statistical targets are obtained from previous test forms to infer nominal parallelism over time.

From a data management perspective, there are three types of information required for test-form assembly: (i) the eligible item pool containing all relevant statistical fields, content fields, and other attributes used in item selection; (ii) a set of constraints specifying restrictions on content coverage as well as other salient attributes (e.g., cognitive skills coding, testing time, reading load indicators, item-parent coding or enemy lists); and (iii) statistical targets. The eligible pool is a pre-screened subset of the

entire master item database. Items are excluded based on prior usage policies, content considerations, or when their associated item statistics signal clear problems (e.g., negative item-total test correlations).

Pearson's test construction procedures are somewhat standard for the educational testing industry, although many testing groups are now moving toward automated test assembly and test-form composition. Some recommended improvements are covered in Section 3.5.6.

3.5.2 Preparation of Scoring Tables

The preparation of scoring tables is a facilitated by the use of the Rasch IRT model for the EOCT and GHSGT. Under the Rasch model, a one-to-one (isomorphic) relationship exists between the number-correct scores and the underlying IRT measurement scale. The actual scaling is and production of the score tables is done by numerically finding a proficiency score on the IRT scale that corresponding to each number-correct score for any set of items. This process generates a number-correct look-up scoring table using only the IRT item statistics for that particular set of items. Scoring tables can therefore be generated for an entire test form or for content-domain-based subscores. The IRT proficiency scale values, in term are linearly transformed to the reported scale scores for the EOCT and GHSGT, with rounding and truncation (of extreme scores—called "lowest obtainable" and "highest obtainable scale scores", LOSS and HOSS) applied as a matter of Georgia DOE score reporting policy.

Using number-correct score look-up tables is very efficient from a processing perspective, because it requires only that Pearson: (a) locate the correct test form look-up table for each examinee; (b) add up the total test points; and (c) look-up the corresponding IRT proficiency score and scale score in the table[14]. From a quality control perspective, it is only necessary to get these three pieces of data correct. In that regard and barring inappropriate software changes in the maximum likelihood scoring algorithms or in the score look-up routines, the principal challenge is data management. We cover that score-table data management issues and quality control in Section 3.5.6.

3.*5.3 Analysis Data File Preparation*

[14] Having to apply IRT scoring to every examinee is highly problematic from a quality control perspective because of idiosyncratic estimation issues and item management issues—especially if more complex IRT models are used. It is possible to use "sum scores" to mimic number-correct score tables in those contexts, as well.

Most psychometric analysis software assumes a de-normalized "*person-by-item*" flat file structure for the data, where item responses (scored or raw responses) are presented in fixed column positions, forming a response vector (record) and each row represents one examinee response[15,16]. Multiple test forms can be represented in a single flat file by creating a master listing of all items used on any of the test forms—possibly sorted by item ID number or some other convenient scheme. The raw or scored response vector is then constructed so that the position of each item in the master listing has a one-to-one correspondence with the serial position of that item in the response vector. Pearson refers to this type of file as an "incomplete data matrix" (IDM) because items in the master listing that are not seen by a particular examinee are represented as blanks or some other missing value in the response record.

The integrity of the IA or other psychometric analyses obviously depends on the integrity of the inferred link between item positions in the master item listing and the column positions in the IDM. Seemingly simple errors such as getting the starting position of the item responses in the response vector wrong (e.g., columns actually begin in column 34, but are inadvertently specified in an analysis as starting in column 33, which coincidentally contains a dichotomous demographic indicator), or sorting the master listing of item IDs after generating the response vectors, could have disastrous consequences.

3.5.4 Item Analyses

An item analysis (IA) serves two purposes. First, it helps evaluate the psychometric quality of each item—primarily flagging items of inappropriate difficulty (too easy or too hard for the target student population) or have little or aberrant sensitivity to the underlying measurement construct, as indicated by correlations between a vector of individual item responses and a corresponding vector of total-test scores. Second, for selected-response items, a distractor analysis helps confirm that the current answer key is indeed the "best answer" from a psychometric perspective. Miskeyed items are

[15] The required de-normalized flat-file structure is largely a convenience mandated by the analysis software, some of which still uses FORTRAN or other archaic programming languages. Older mini and mainframe systems typically had 80 or 132 character record length limits, forcing longer response vectors for a single person to be recorded on multiple records. Those limitations have largely disappeared for modern computer systems, but the required flat-file structure for data records remains.

[16] Many of newer statistical software packages provide the capability to read delimited files (e.g., comma-delimited, tab-delimited, or some other character). Delimited files, however, require a parser. Unfortunately, different parsers—especially when applied to files created under different operating systems or from certain applications—may result in difficult-to-detect data read errors. Fixed column flat files, while not very efficient from a file-size perspective, are usually less prone to data read errors and easier to check by visual inspection.

usually detected in the IA by a negative correlation between the item responses and total test scores. The "correct" key or a secondary key is typically indicated by a positive correlation between responses that endorse that option and the total test scores, implying that the highest proficiency examinees in terms of total-test performance also tend to choose that particular response.

The IA carried out by Pearson for the EOCT and GHSGT test forms appears to be fairly standard for the testing industry. Per the Georgia functional requirements, items are flagged during the IA if they meet any of the following conditions: (a) mean item proportion-correct of $p < .15$ (difficult items near a chance guessing level of difficulty); (b) scored item response-total score correlation, $r_{pbis} < .20$; and (c) incorrect distractor option selected by 40 percent or more students[17]. Pearson software item analysis application is referred to as the "TRIAN process" in their documentation. The process appears to be run by the Pearson information technology group. Inputs are raw student data (unscored) and keys. Output includes classical item statistics, stored in a flat file. Answer keys are reviewed for all flagged items. Any key changes are made in the test map file and entered into ChangeMan™, a change tracking device used by the information technology group. A software testing group verifies that all changes have been made.

The approach of logging key changes in ChangeMan and subsequently uploading those changes to the master item database does not imply that the master item database serves as the source record for all subsequent psychometric analyses. We address this potential issue in our recommendations (Section 3.5.6).

A final item analysis is carried out *post*-administration. Those finalized IA statistics are assumed to be loaded to the master item database via Pearson's Item Tracker application and all changes verified.

3.5.5 IRT Calibration and Equating Steps

Because of the very tight turn-around time required between the end-of-testing dates and score reporting, Pearson and the Georgia DOE—with the endorsement of the Georgia DOE Technical Advisory Committee—use a pre-equating model for scoring current EOCT or GHSGT test forms. This pre-equating model exploits a one-to-one relationship between number-correct scores and IRT proficiency scores that are

[17] Many testing programs supplement this "popular incorrect key" flag with a flag for incorrect response endorsement that correlate positively with the total test scores. This latter type of flag can help identify possible secondary keys (or the correct key in the case of a miskeyed item) that are correctly selected by the higher performing students. See our recommendation in Section 5.3.2.

estimated to directly take into account differences in test-form difficulty. A scoring table is generated for each test form, using IRT item statistics that were calibrated and equated to a common metric for each examination, based on either prior field-test or operational use of the items. These scoring tables can be prepared in advance of the actual test administration, based on the finalized test-map files. Early return samples from the current test administration are used to check the calibrated IRT item statistics used to generate the scoring tables. In the event of serious discrepancies, scoring could be delayed until all items are re-calibrated and equated using the Georgia census samples for each test.

As already noted, the primary source information and data for every test form is contained in the test map resource file. Prior to pre-equating, the test map files need to be verified against the master item bank. Pearson manually carries out this step, checking the items in each test map file for domain, content standard, answer key, IRT Rasch item difficulty estimate, and depth of knowledge coding. In aggregate, by test form, manual checks are also carried out for item counts by status (operational versus field test), by content domain, and in terms of shared (common) items between the two core forms of each test.

The pre-equating steps are independently carried out by two psychometricians at Pearson with required 100 percent reconciliation and matching of both sets of results. The pre-equating is done by pulling IRT item statistics for each test form (test map). The item statistics are then used to numerically find the IRT proficiency score associated with all-possible number-correct scores on that test form. The proficiency scores are then linearly transformed to the reported scale scores. This makes it possible to generate a scoring table for each test map (form), using only the associated item statistics. The key becomes the identification number associated with each score table. Each table must be a sole-source data object that is infallibly linked to all examinee records by the same identification number. If the links become corrupted in any way, incorrect scores will result for all examinees using an inappropriate scoring table.

As noted above, Pearson uses a "double entry" method of quality control, and borrowed from the accounting field, to assure that these types of problems do not occur. That is, two psychometricians independently run the analyses and are required to exactly match their results. If they do not match, they reconcile the results to detect and correct the errors so that they do match[18]. The two psychometricians employ various

[18] We assume that the reconciliation results are logged into a database of record. If not, we recommend that this type of log be created as full documentation of the processing steps.

policy-dictated exclusion rules (i.e., total raw score is missing or zero, blank response vectors, duplicate records, acceptable form numbers not matched). However, the Pearson documentation does not provide details regarding one possible loophole concerning the integrity of the source files. That is, if both psychometricians are using the same interim files (master item file and IDM) as the source of their analyses, they could fully agree on results based on the same "bad" data. We return to this issue by way of our recommendations in Section 3.5.6.

The actual calibration and equating process is carried out in three stages. The first stage involves a "local IRT calibration" of the IDM for a particular domain, using the WinSteps™ software package. The second "equating" stage involves an iterative stability analysis involving only the reused "anchor" items—that is, items that already have IRT statistics calibrated to the base metric. The stability analysis computes $b_i^*=b_i+d$ for the anchor items, where $d= \mu(b_{base.scale})- \mu(b_{local.calibr.})$, excluding at each iteration those items whose difficulty estimates on the base scale and adjusted difficulty estimates, b_i^*, differ in absolute value by more than a predetermined "displacement"—typically taken to be .3. In the final stage, the anchor items that survive the stability analysis are used to fix the calibration scale by constraining their item difficulties to be equal to the base scale difficulties. This is oftentimes called an "anchor calibration".

There are complications associated with carrying out an anchor calibration. First, the proper item difficulties for the anchor items must be downloaded from the master item database—the master "item bank"[19]. Second, the serial listing of unique item identifiers, and the columns of the incomplete, scored data matrix (IDM) must be synchronized so that the proper item statistics are associated with the proper item identifiers. Finally, the anchor-item difficulty estimates must be matched to the serial position of the items in the calibration list (and to the corresponding columns of the IDM) for the final "anchor calibration". Any discrepancies should generate flags that require resolution before processing can be finalized.

[19] Various organizations have different policies about item banking and how the "base scale" is maintained via the item difficulty estimates stored in the "bank". Some organizations use only the most recent, equated estimates of the item statistics as the base-scale values. Others create a weighted average of the previous and new [equated] estimates as the best linear. If the weights are proportional to the error variances of estimate, the weighted averages are optimal and unbiased. Still other organizations essentially recalibrate all of the data across years and reset the base scale from the multi-year data. The Georgia EOCT and GHSGT programs use the "most recent" approach to banking the item statistics.

3.5.6 Recommendations for Psychometric Data Management and Processing

Although psychometric analyses are usually quite technical, most of the quality control steps for the analyses can be handled by: (a) confirming the integrity of the queries and intermediate analysis source files generated by the queries; (b) confirming control parameters, options, and file names in the analysis set up files; and (c) evaluating critical outcomes in line with expectations. The first two steps largely involve data management, but should not be treated as *de facto* correct, just because a query or extraction ran without reported errors. Counts need to be match exactly at multiple levels (e.g., numbers of operational and pretest items for each test form, numbers of examinees taking each item). Pearson does require two psychometricians to independently carry out most of the critical analyses and reconcile their calibration and equating results to exactly match each other. It is less clear that the queries and data extractions to generate the analysis files go through similar reconciliation steps. For example, the queries of critical item information such as answer keys or IRT anchor item statistics pulled from the master item bank, or the extractions of examinee-by-item record used to create the IDMs needed for the item analyses or calibrations and equating steps should be independently carried out with separate queries by two analysts and confirmed to 100 percent match insofar as the integrity of the data. This same "two eyes" (2Is) principle needs to apply to generating scoring runs as well. It is a simple human mistake to fail to change a scoring table naming reference in a script. But the resulting score-reporting errors can be disastrous. These types of routine operations cannot be delegated to one database manager or an analyst, even if a query application and interface are constructed to facilitate the processing. The 2Is principal needs to be applied to <u>every</u> level of the processing, especially when procedures are manually initiated or run (i.e., 2Is for all set-ups, queries, and source file extractions).

Another useful principal is to "reconcile to expectation" (R2E) insofar as expected gross record counts, missing data counts, numbers of items seen per examinee (minimums and maximums), etc.. Even equating results can have expected outcomes. For example, if we knowingly increase the difficulty of the test forms—perhaps by moving the measurement precision targets to improve key decisions—R2E would imply that we expect a positive equating constant from the equating stability analysis (Section 3.5.5). If we instead obtain a negative equating constant, we need to understand the discrepancy from our expectations (e.g., item drift or some other plausible psychometric explanation). Equating results and a discussion of any discrepancies should become part of the test map record.

3.5.6.1 Specific Recommendations for Test Construction

Like any manufacturing process, test assembly and composition should seek outcomes that are consistent with specifications. Sometimes, in testing, the test specifications, themselves, are somewhat ambiguous or suboptimal on psychometric grounds. In that regard, it is possible to perfectly match somewhat flawed specifications and still wind up with an inferior test. The EOCT and GHSGT specifications should include, at a minimum, two types of information: (1) fixed statistical IRT measurement information targets that specify maximum measurement precision at key decision points on the corresponding scales, and (2) fixed content specifications including fixed counts or ranges of acceptable content frequency. In addition, tolerances and violations of the content specifications and/or attainment of statistical specifications should be carefully documented.

These fixed specifications should be stored and used for assembling each test form. Out-of-tolerance forms should be carefully scrutinized and authorized for use if and only if all reasonable attempts to locate usable items fail. And then, the forms should be treated as exceptions, not used as the basis for building the next round of new test forms. The current approach to test construction appears to be based on an evolving set of test specifications, where each previous set of test forms is used as the as the basis for building the next set of forms. From an engineering perspective, this approach is essentially one of propagating the errors in test construction by modifying the specifications to match the outcome. There is a better way. If, for example, content standards change, or if the Georgia Technical Advisory Committee recommends changing the measurement precision targets to improve precision at a key point of the scale, those types of changes should be handled by modifying the specifications.

The specifications should also guide the item writing and "inventory replenishment" over time. Most violations in meeting test specifications occur because of insufficient supply of items to meet the demand. A well-designed inventory control system anticipates shortages in key content areas or within key item difficulty ranges, and attempts to fill those voids in the item bank before they reach a point of critical shortage.

Variations from the targets can also be used to signal likely changes in the score distributions and help justify greater-than-normal equating adjustments. For example, if the test forms are made more difficult—by design or by virtue of the increased average difficulty available items in the pool—it is possible to anticipate a positive equating constant (see Section 3.5.4). Of course, changes to the formal specifications

require solid psychometric justification and addition communications between relevant parties.

3.5.6.2 Recommendations for Preparing and Managing Analysis Files

Data files should be verified and reconciled to exact counts. Create the following file object relations: (i) test forms (contains a master listing of test map identifiers); (ii) a master listing of examinees by test maps (i.e., examinee identifier and test map identifier); (iii) a master listing of items and test maps (that is, test map identifier plus item identifier for all items); (iv) a master listing of all items with pretest or operational status indicators—extracted from the master item data base; (v) a master listing of item sets (set identifier and item count, where discrete items have a set length of one); (vi) a listing of items by item sets; (vii) a listing of examinees by item transactions (i.e., examinee identifier, item sequence, timing data if applicable, as well as raw and scored item responses). The R2E principle applies here, where every item, every examinee record, and every response record is counted and used to verify the expected numbers of test forms, item sets, items, and examinees.

Reconcile to expectation (R2E) the item response records for all examinees. Account separately for omits versus not presented items. Flag and check corrupted records, prior to the analysis. Reconcile examinee counts by items (unused items or non-equating items seen by all examinees). Reconcile items not appearing on any test form. Reconcile examinee test map identifiers not in the master list. A data source integrity sign-off should contain a prescribed quality control protocol. Automated flags can be implemented to prevent processing until discrepancies are resolved. Or, at the very least, a series of warnings may be broadcast to the TMT, information technology team, and to the psychometric team for action and follow-up.

Whenever incomplete data matrices (IDMs) or other analysis files are created, store the query and an image of the extract and analysis data file. Do not just store the query. A dynamic data system may change causing a different interim query output, even from the SAME query executed at an earlier time.

Any changes in the item answer keys or other changes in primary data source for test forms, items, or examinees should in the master database. For example, if there is a change in any answer key, that change needs to be made in the master item database. Rerun the extract, reconciliation steps, and then and only then the psychometric analyses of interest. Changing the answer keys or other data only in the interim listings creates "dual source" for the keys and can lead to problems.

3.5.6.3. Recommendations for Item Analyses

We have two recommendations for item analyses (IA). One recommendation is that IA-generated statistical flags be uploaded to the master item data base, at least following the final IA. A second recommendation is to add a new flag for incorrect distractor endorsement showing a distinctly positive product-moment correlation, $r(d_{ij},X)>.05$. It is acknowledged that this type of change would require a modification to the EOCT and GHSGT functional requirements and has contract implications.

3.5.6.4 Recommendations for Producing Scoring Tables

The preparation and 100 percent reconciliation and matching underlying the production of the score tables by two independent psychometrician directly implies the 2Is principle and provides excellent quality assurance of the score table, itself.

3.5.6.5 Recommendations for Calibrations and Equating

The calibration and equating procedures used for the EOCT and GHSGT are relatively straight-forward. Furthermore, Pearson has two psychometricians independently carry out the analysis steps and fully reconcile their results. Pearson also uses a well-established program, *WinSteps*™, for the calibrations. *WinSteps* is used by many testing agencies because: (a) it handles extremely large data sets; (b) deals easily with different mixtures of item types; (c) readily handles multiple test forms and incomplete calibration designs.

The equating steps (local calibrations, iterative stability analyses, and final anchor calibrations) are also fairly straightforward. One recommendation is to generate summary reports that ensure that the Rasch anchor item difficulties have been properly pulled and been recorded in the item anchor file along with the corresponding item sequence positions in the IDM as part of the *WinSteps* anchor-item calibration. Because of the way that *WinSteps* centers the scale during the anchor calibration, merely checking the resulting mean of the item difficulties does not guarantee that the proper items were necessarily anchored.

3.5.6.5 Recommendations for Score Reporting

The use of scoring look-up tables is an extremely efficient way to implement IRT-based scoring, without the complexity of individually estimating IRT maximum likelihood estimates for every examinee. As described earlier, instead, a single number-correct score look-up table is generated for each test form. The look-up table includes the number-correct scores and corresponding scale scores for each test form. The scale

scores are computed from the equated IRT item statistics for each test form. Scoring therefore has three simple steps. First, find the proper scoring table for each examinee's test form. Second, compute the examinee's number correct score. And third, look up the corresponding scale score in the scoring table. This scoring process is so simple that it would seem to be almost fool proof. That is not necessarily the case.

A potential problem occurs if the name of the scoring table is not explicitly confirmed and rechecked as part of the test map, and locked down so that in incorrect table (e.g., an earlier version) can never be called for final scoring. This is especially problematic if application software scripts are used to implement the scoring. In those instances, the script should be manually composed—at least insofar as adding filenames, table names and other critical parameters—to ensure that the proper data are incorporated in the analysis. A better approach is to establish naming conventions for the scoring tables beforehand and enter each score table file name as part of the test map specification of record. Scripts can be used to implement the look-up process (scoring), however, the score table file name(s) entered into the scripts should be automatically generated from the test map file or manually entered by the systems operator or psychometrician in charge of running the scoring application, and then double-checked by an independent process. Scoring table file names should never be coded into processing scripts by a single analyst because a simple clerical error such as failing to change the file name can have disastrous consequences. There is also a tendency to reuse scripts over time, without proper versioning controls in place to ensure that the critical parameters—such as scoring table file names, source file names, etc.—have been properly changed in the script that is actually run. Having 2Is at least review <u>every</u> script and query is also good practice—especially for routine but critical processing steps.

A reasonable, follow-up quality control check is to always generate the score distribution mean and standard deviation for a defined early-return sample. Those outcomes should always be checked against expectations, verified as being different from recent, prior administrations to ensure that earlier versions of the scoring tables or an improper source files were not inadvertently used (e.g.., confirming that an older or improper table or filename was not specified in a script), and then double signed off. The table-based scoring summary results should also be rechecked against linearly transformed mean and standard deviations of the IRT proficiency estimates, at least for an audit sample of test forms and/or examinees. Scaled proficiency estimates can be readily obtained by anchoring in the WinSteps calibration software all operational item parameter estimates for each form at the same values used in producing the score

tables. If the two sets of results—that is, the moments of the score table-based lookup scores for the examinees completing a particular test form and the moments of the transformed IRT proficiency score estimates for the same examinees—do not match (within acceptable rounding tolerances), the specific score tables with possible problems can be isolated.

4.0 Review of Past, Specific Issues for the EOCT and GHSGT

Obviously, the best approach to dealing with problems is to implement proactive, preventative measures. Yet, despite the best strategies, unforeseen problems occasionally occur. In those cases, it is important to understand the cause and to ensure that appropriate, preventative procedures have been implemented to absolutely ensure that similar problems will not occur in the future. This section outlines five specific issues that have arisen over the past year or so involving Pearson's work on the Georgia EOCT and GHSGT. It is unfortunate that these problems occurred and further unfortunate that each new incident chips away at the foundational credibility and trust that the Georgia DOE has in its vendor.

Pearson has implemented a process called "root cause analysis" (RCA). Each RCA is a documented process that describes the nature of the problem, its impact and numbers of examinees affected, date of discovery, discovery source, root cause, proposed solution(s)/action plans, and implementation of the solution. A RCA was carried out for all five issues.

While quality-assurance procedures such as root cause analyses are useful, they remain reactions to specific problems or issues and may not address more systemic issues. In that regard, our review of these four issues consider: (a) whether the problems might have been preventable had some of the recommendations in this report been in place and (b) whether the problems might be part of larger, systemic issues.

4.1 Wrong Extract for Winter 2008 Scoring

This issue involved the EOCT Economics and 9th Grade Literature examinations. 7,462 had incorrect scores reported on an electronic report produced in December 2008. The paper version of the report was accurate. The reporting error was traced to the use of the wrong extract file in preparing the electronic report. A Production Readiness Review, to be conducted by Pearson program management, has been recommended as a fix. The implementation date for the fix is 30-October 2009.

This issue seems entirely preventable. There were two *report forms*, one producing an electronic report and the second producing a paper report, and each was generated from a separate extract. Throughout this report, we have emphasized the need for single-source files (master database files, extracts or query-based tables). In this particular case, a single extract should generate both reports. Multiple extracts can and should be independently generated to verify the integrity of the query. Additional reconciliation steps should further be taken to confirm that the extract is correct. Once verified, however, any extract needs to be locked down and used as a single data source for all reports. It appears that the current process may still allow multiple extracts to be generated for the various reports and the Production Readiness Review used to [hopefully] catch and deal with discrepancies. A locked-down, verified extract, with independent checks and double sign off (2Is), would be a better solution. Similarly, the report forms themselves, should be double-verified and locked down to prevent the wrong report form from being used (e.g., a report with inappropriate headers, sort conditions, summary statistics).

It is not clear whether the current or improved procedures use scripts or database structures to define the source files (extract, query, etc.), report form(s), and report generation parameters to use. The latter, database approach would obviously be easier to verify, provided a better documentation of every report generated, and be less prone to human errors (e.g., failing to change a filename or reusing an old script with improper settings).

4.2 Math Formula Sheet Issue

This issue involved a versioning problem for the Math I and Math II EOCTs and was discovered in June, 2009. A correct formula sheet was pulled for the sample items, but was not the correct sheet for the operational items. 396 Math I and II students were affected.

The RCA identified as this as a exception processing issue. The correct process is to replace the test resource [source] images, rather than uploading new files to multiple instantiations of each test form. The resolution is to define two processes for replacing images: one for complete replacement of test images and the other allowing multiple images to exist, if allowed by program requirements.

This issue seems to be preventable by: (a) defining a template for each test map and attaching formula sheets, etc. to the template; (b) referencing the template in each test map; and (c) creating a unique instance of each test map (form) if any changes are made late in the production cycle (e.g., map XYZ.1 versus XYX.0). As discussed rather

emphatically in Section 3.1.1 and 3.1.2, each item and each test map should only exist in one and only one place. That holds for the resource data, answer keys, and images. The best solution—especially if it occurs late in production—is to create a special, new instance of the subject data object, one with a unique identity that can be specifically tracked as an exception instance—possibly with parent reference. Two or more data objects or images with the same identifiers should simply not be allowed to exist…ever. The "two process" resolution alluded to by Pearson seems to potentially violate this rather straightforward principle.

4.3 EOCT Scoring Issue

This issue apparently occurred in April 2009 and impacted 407 students who took online EOCT tests. Students who viewed and skipped two or more consecutive questions were scored incorrectly, due to improper handling of blank responses. For example, if a student answered items 1-12, left items 13-15 blank, and then answered items 16-20, the results record showed that the student attempted items 1-12, skipped item 13, attempted items 14-18, and did not attempt items 19 or 20.

The RCA identified an incorrect version of the response record processing code, due to improper version control. Better version control with required test cases, sign-off, and lock-down will supposedly address this issue in the future.

To the extent that the problem is exclusively related to programming version control, it appears that this particular problem has been resolved. Pearson's RCA specifically mentions that a new Perforce version control application has been implemented. It is essential, however, that all new version changes be evaluated extensively using a full range of test cases, with multiple analysts reviewing the results, and making it virtually impossible for a single operator to implement changes in an operational system without going through all required QC steps and sign-offs.

4.4 Survey Item Issue

This issue was, on the surface, seemingly rather minor and only involved an unscored survey items. However, it may have broader implications. The problem occurred in March 2009 and involved that a survey item was set as item category = "Survey". Unfortunately, "Survey" is not a recognized in the subsequent scoring extract utility as an answerable item type and all of the survey responses were excluded. Ordinarily, the test map team loads survey items into authoring and then change the item category field to "Field Test". There are basic QC processes are designed catch invalid responses

(i.e., "true" errors) for answerable items. In this case, Pearson's RCA stated that a "false" error is created and did not show up on any QC reports.

We have three recommendations regarding this issue that extend beyond this incidental root cause. First, policies about the status of individual items (or any associated data objects) should reside with test map. Each test map should undergo a complete reconciliation with summary reports that show counts complete listings of all scored items, unscored items, and unidentified items. That type of report should be able to flag any aberrant coding of data fields. Second, all data fields for well-established entities such as item types should reside in a secondary, normalized look-up table, to prevent anybody from simply plugging in a convenient field name or other arbitrary value. That is probably not what happened here, but forcing a look-up would prevent "Survey" from even being entered; rather than requiring an *ad hoc* change at publication time. Item types change rarely enough that there is no reason to manually enter associated item type codes for each item anywhere but in a unique item-type table. Third, all queries and data extracts need to work from a list of test maps for purposes of generating analysis files and other extract files used in processing. Each test map in the list should be the primary data source extract containing the essential test map data (e.g., item list, sequencing, answer keys pulled from the master item database). Fourth, any primary scoring query/extract routines need to include a complete check and reporting of counts showing, for example, complete records, listings of partial records, items not seen by any examinees, test forms not match to examinees, examinees not matched to test forms, unknown items (e.g., a "survey" code), etc..

4.5 Score Extract Upload Error

It was discovered that the incorrect district test score files were uploaded to the Georgia DOE secure FTP site in August 2009. The files included results from the Spring 2009 data, rather than the proper results for the Summer 2009 test takers. An independent quality assurance verification of district files was not performed by Pearson after files were generated. This was a violation of the Pearson's own internal policies. The correct extract was generated and available, but not uploaded. Remedial action by Pearson includes anticipated implementation of procedures that specifically preclude either accidental or deliberate execution of un-validated software programs to generate customer deliverables, better training and documentation, and tighter production software access and execution controls. The documentation, training, and implementation of these procedures will be in place by November 2009 and will further

be tracked and verified Pearson's quality assurance group. Compliance will be audited periodically by the quality assurance group.

It appears that Pearson's quality assurance steps will help prevent a recurrence of this type of problem, if the proposed improvements—especially the executive software policy controls preventing the use of un-validated software, queries, or other extract utilities used to generate results of any type for the Georgia DOE—are ubiquitously implemented. We also recommend an addition step.

A summary database should be created to keep track of all final, verified versions of all analysis files, results files (IDMs, scores, etc.). This master file-listing database would become the primary source for pulling file names or table names for uploads, etc.. The master listing of the data file names (including relevant counts and file content descriptions) would allow Pearson staff to easily check for possible multiple versions, as well as providing easy access to what has been created. Sign-offs (verification) of the files should be readily apparent in this master file listing. Forcing all software extract or other production routines to use the master file listing would further ensure that the most recent version of each file is being employed and provide a direct way to implement and control the executive policies noted above.

5.0 Overall Evaluation

Constructing, administering, and processing test data for large-scale examination programs like the EOCT and GHSGT requires a very complex enterprise and coordination of enormous resources. In general, Pearson demonstrated an impressive array of systems and procedures that seemed more than adequate to handle the work. Personnel at the Iowa City facility also appeared to be highly competent and cooperative insofar as providing explanations and information about their individual roles and operations.

However, that five examination processing incidents occurred since December 2008 is extremely troubling. Considered cumulatively, the five processing incidents either suggest a run of "bad luck" or possibly systemic problems. All five of the incidents were traced to problems of human errors, failures to check the data, or exceptions that were not properly handled. Systemically, it could be argued that real causes were: (a) overly loose quality assurance regarding version controls and/or too much latitude by individuals to make on-the-fly changes to scripts, software queries, and reporting applications; (b) insufficient verification of set-ups and manually initiated procedures; and (c) incomplete verification and sign-off of results. This DMR also

suggested possible future loopholes, including: (i) failure to routinely and completely reconcile key data sources (i.e., master database files) against extracts and other "detached" data files (e.g., test map resources); (ii) passive flags, logs and exception reports that possibly allow certain late-cycle exceptions to propagate as errors elsewhere in the system; and (iii) and insufficient lock-down of certain critical files and/or directories. Well designed systems should minimize human errors, require multiple checks and sign-offs for all human-initiated activities, curtail exceptions, and treat exceptions in a very serious way with prescribed resolution and additional quality control steps when they do occur. If systems are being continually repaired to manage exceptions, the entire system design may require serious reconsideration. Pearson's systems are good and do not appear to require complete redesign; but, those systems are far from perfect and can be improved.

Overall, we give Pearson a "B", implying that there is room for improvement. We have four general recommendations that we believe would move Pearson's grade to an "A".

Our first recommendation is that Pearson develops a larger array of automated mechanisms—that is, software applications and routines—for reconciling every aspect of the data throughout the processing cycles. *The goal here is to comprehensively check for discrepancies*. We believe that most discrepancies procedures and reports can and should be fully automated. Examples include comprehensive comparisons among test forms or between test forms and master databases (e.g., comparing images on test maps to the master item data base, comparing keys across test maps and against the master item database, reconciling all unknown entities such as "survey" items to a master listing of active content). Discrepancies should not be passively reported and then possibly ignored[20]. Rather, all discrepancies—depending on their severity as determined by predetermined policies—should always generate *automated* flags that: (a) stop further production or at least activate a series of prescribed steps (e.g., activate an exception handling agent); (b) require resolution with active sign-off; and (c) mandate secondary follow-up checking to ensure that the resolution and sign-offs were implemented.

[20] Many of Pearson's exception logs and reports appear to be entered into Microsoft Word or Excel documents. Although that is fine for human review, it is not clear that these files are linked to requisite actions or that automatically halt further processing until the discrepancies are resolved. For example, an item key validation discrepancy flag should requires concrete resolution and action to remove the flag—not just passive human review of an Excel file and [possible] sign-off—before the item analysis or scoring can proceed.

Our second recommendation is that Pearson make every attempt to implement procedures that require changes to be directly performed in the master database sources (e.g., changing answer keys, item images, etc.), rather than allowing modifications to extracts and or "detached" versions of the data records. Although the test map is a wonderful concept on the surface, providing strong controls over every test form, it seems possible that different versions of the same items can presently exist on different test maps under the same identifier. Although the test map resource data is extensively reviewed by the test map team, editors, and others, it is not and never will be the primary source for the items. Mathematics items can likewise be tagged to formula sheets or other resources (e.g., activating a calculator agent on a computerized test form). This would preclude errors in generating formula sheets or other exhibits at the test-map level. Ultimately, steps should be taken to always change the data at the primary source! Where it is simply not possible (e.g., due to late-cycle cost constraints related to reprinting massive quantities of test booklets), extensive *automated* reconciliations and discrepancy resolution activities should be required as noted above, with guaranteed "final" changes resolved in the master database.

Our third recommendation is to move to database representations of as much of the processing as possible, as opposed to using scripts, SQL code with hard-coded file or table names, etc.. For example, scripts or program code should use tokens for file names and other set-ups. The tokens should point to fields in database records. Each of the database entries for the query or analysis can and should undergo 2Is (see Section 3.5.6). It appears that much of the end processing of reports and score extracts is currently managed by queries or other scripts being run by possibly only one operator. A database approach provides better documentation and makes it easier to carry out strong quality control procedures, especially if the query, analysis or follow-up routines automatically run reconciliations and upload discrepancy summary results to the query/analysis set-up database.

Our final recommendation recognizes that it is virtually impossible to fully automate any system. That being said, *where ever humans touch the systems or data (e.g., manually preparing scripts, running extract queries/SQL, or entering field names in analysis records in a database),implement mandatory 2Is with active sign-offs, as noted in Section 3.5.6*. Also, all results, from equating to score reporting should employ the reconcile-to-expectation (R2E) principal, again, with active sign-off. The test map team (TMT) currently does take responsibility for many checks, the psychometrics group uses a dual processing approach for the calibration and equating steps, and other groups such as the materials handling group require 100 percent audits of test materials. However,

there are still some apparent loopholes in the system, as evidenced by the incidents described in Section 4.0.

We acknowledge that this Data Management Review reflects possible limitations on our understanding of some aspects of Pearson's systems operations. No audit is ever infallible. In the spirit of cooperation and towards the ultimate goal of improvement and error-free operations, we hope that any such misunderstandings can be resolved in subsequent phases of the PPSA.

Appendix A

CART Audit of Pearson

Information Request

09 September 2009

The Center for Assessment and Research Technology (CART) has been authorized to carry out an audit of test development and assessment processing systems and procedures at Pearson relative to select examinations for the Georgia Department of Education. CART is requesting documentation for the Data Management Review phase of the Audit and relevant to the specific points outlined below under bullet #1.0. An on-site visit to Pearson is currently scheduled for 23-24 September 2009. A suggested list of topics for that on-site visit is covered under bullet #2.0.

1. Documentation
    - 1.1. Understanding of Georgia DOE's assessment and processing needs
        - 1.1.1. Performance indicators
        - 1.1.2. Identifying processes, objectives and operations
        - 1.1.3. Understanding the scope of test processing for the state of Georgia
    - 1.2. Site locations and operations (including distributed operations)
        - 1.2.1. Operations and processes (manual, semi-automated/human-assisted, fully automated)
            - 1.2.1.1. Test development
                - 1.2.1.1.1. Item authoring and coding
                - 1.2.1.1.2. Answer keys and rubrics: production and management
                - 1.2.1.1.3. Modifications to item-level fields
                - 1.2.1.1.4. Management of item sets
                - 1.2.1.1.5. Item analysis and key validation
                - 1.2.1.1.6. Test assembly and test form integrity
                - 1.2.1.1.7. Packaging and deployment of test forms
                - 1.2.1.1.8. Quality assurance and quality control procedures
            - 1.2.1.2. Test administration
                - 1.2.1.2.1. Managing test eligibility issues
                - 1.2.1.2.2. Forms assignment/spiraling
                - 1.2.1.2.3. Security management for all test materials
                - 1.2.1.2.4. Transport/transmittal of answer sheets/response data
                - 1.2.1.2.5. Exception management/troubleshooting

1.2.1.3. Initial processing of completed tests
1.2.1.3.1. Reconciliation of test forms and examinee data
1.2.1.3.2. Handling of aberrant answer sheets/response records
1.2.1.3.3. Preparation and cleaning of raw response data
1.2.1.3.4. Preparation of item analysis files
1.2.1.4. Item analysis and key validation
1.2.1.4.1. Software and detailed IA procedures
1.2.1.4.2. Key validation and management of IA-flagged items
1.2.1.4.3. Management of items with answer-key changes
1.2.1.4.4. Production of new data files with key changes made
1.2.1.4.5. Final item analysis and management of IA item statistical results
1.2.1.5. Calibration and equating steps
1.2.1.5.1. Production of final scored data files for calibration
1.2.1.5.2. Calibration steps and evaluation of results
1.2.1.5.3. Equating steps
1.2.1.5.4. Anchor calibration steps and QC
1.2.1.5.5. Item banking of equated IRT statistics
1.2.1.6. Scoring and reporting
1.2.1.6.1. Item-level response scoring and QC
1.2.1.6.2. Production steps for scoring tables for test forms
1.2.1.6.3. Examinee-level total-scoring procedures
1.2.1.6.4. Aggregation of scores by units of interest
1.2.1.6.5. Reporting steps and QC of results
1.2.2. Information systems and data management
1.2.2.1. Data management: primary and secondary item repositories
1.2.2.1.1. Item presentation content
1.2.2.1.2. Generation of files for test assembly (item content and statistics)
1.2.2.1.3. Storage and management of answer keys and rubrics
1.2.2.1.4. Statistics used in scoring (e.g., equated IRT difficulties)
1.2.2.2. Test-form data management
1.2.2.2.1. Test composition and forms management
1.2.2.2.2. Test assembly constraints and targets
1.2.2.2.3. Test form identification, printing, spiraling and packaging
1.2.2.2.4. Test booklet recovery and reconciliation
1.2.2.3. Data management for repository for examinee data

1.2.2.3.1. Examinee records and storage
1.2.2.3.2. Answer sheet/record recovery and reconciliation
1.2.2.3.3. Production of analysis and scoring source files

1.3. Allocation of resources
1.3.1. Operations
1.3.2. QA/QC and internal evaluation (metrics and variables)
1.3.3. Effectiveness (metrics and variables)
1.3.4. Improvements (proactive vs. reactive)

2. Onsite discussions at Pearson (Iowa City)
2.1. Interviews/meetings with personnel and operations
2.1.1. Item editing and data entry
2.1.2. Test development, assembly, composition and packaging
2.1.3. Test administration and operations
2.1.4. Scanning, reconciliation and initial data cleaning
2.1.5. Psychometric processing, scoring, and reporting
2.2. Data management
2.2.1. Manual procedures and hand-offs
2.2.2. Verification and quality control checks
2.2.3. Sign-off
2.2.4. Lock-down controls on verified data

Psychometric Procedures and Systems Audit:
Examination Processing Review
for the Georgia End-of-Course Test (EOCT)
and Georgia High School Graduation Test (GHSGT)

Richard M. Luecht, PhD

Terry Ackerman, PhD

Center for Assessment and Research Technology (CART)

Greensboro, North Carolina

March, 2010

1.0 Introduction

The CART Psychometric Procedures and Systems Audit (PPSA) is a three-phase, technical evaluation of the end-to-end flow of item and examinee data. Overall, the audit evaluates many aspects of the assessment processing system, from test development through scoring and reporting, and requires a high level of operational psychometrics expertise covering all phases of examination processing and advanced computer-based testing systems design and implementation expertise. This report summarizes the Examination Processing Review (EPR) part (Phase 2) of the PPSA carried out on behalf of the Georgia Department of Education, in cooperation with the Educational Measurement Group of Pearson in Iowa City, Iowa. All phases of the PPSA are limited to Pearson operations as they apply to the Georgia End-of-Course Tests (EOCT) and the Georgia High School Graduation Test (GHSGT).

The EPR is an operational, technical evaluation of the results generated by various software programs and manual procedures used in end-to-end processing of an examination. Almost any examination program can be conceptualizes as having two pathways. One pathway involves the assessment tasks or items that are entered into the system along with other information such as answer keys and rubrics, content codes, etc.. The items typically progress through two stages of use, from initial tryout or pretest through operational use (i.e., scored items). The items are selected for test forms, deployed, administered, and ultimately analyzed. At any of these phases, data or data management errors can occur. The purpose of the Examination Processing review is to evaluate the data processing of items and test forms. The second pathway involves the examinees. Examinees enter the testing system via some type of registration process that may include a formal application where eligibility is checked, payment is collected, an examination authorization is issued, and the examination is scheduled. Or, the examinee may simply enter the system upon receipt of a test form answer sheet or receipt of a set of electronic examinee response records (for computer-based examinations). The EPR follows the examinee pathway(s) through the system as well, checking for integrity of the response data, loop holes in the system that could contaminate or corrupt subsequent psychometric analyses or lead to scoring errors.

The EPR focuses on the quality control/quality assurance (QC/QA) steps by reviewing empirical results. The review specifically looks at sample intermediate processing results, including: data reconciliation and cleaning; item analyses and key validation procedures; calibration and equating; scoring; and reporting procedures.

2.0 Scope of the Examination Processing Review

The PPSA covers and evaluation of Pearson's data management systems and processing procedures related to the Georgia End-of-Course Tests (EOCT) and the Georgia High School Graduation Test (GHSGT). Pearson's contractual obligations for test development, administration, and processing are summarized below.

The Georgia EOCT is administered in grades nine through twelve for eight state-mandated core subjects: (i) Mathematics I (Algebra, Geometry, and Statistics); (ii) Mathematics II (Geometry, Algebra II, and Statistics); (iii) U.S. History; (iv) Economics (including Business and Free Enterprise); (v) Biology; (vi) Physical Science; (vii) Ninth Grade Literature and Composition; and (viii) American Literature and Composition. In addition, legacy Algebra I and Geometry test forms are administered to accommodate students who entered high school under the previously authorized Georgia Quality Core Curriculum (QCC)[1]. Any student taking an EOCT course, regardless of grade level, is required to take the corresponding EOCT upon completion of that course. EOCT scores are averaged in as 15% of each student's final course grade. New EOCT tests are usually constructed for winter and spring administrations. Recycled tests are used for the summer administration and mid-month administrations. The EOCT can be administered via paper-and-pencil assessments or in an online format. Paper-and-pencil assessments are only available during the winter, spring or summer administrations. The EOCT items are developed by Measured Progress (MP) in collaboration with the Georgia DOE staff and high school educators. The test forms are designed and developed by Pearson staff, who also take full contractual responsibility for the following activities: (i) providing comprehensive program management, (ii) overseeing item development by MP, (iii) providing psychometric services including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the EOCT program, (iv) creating customized administration procedures for receipt control, data editing, and scoring processes, (v) designing, printing, and distributing all test materials and ancillary documents, including electronic and Braille test versions, (vi) processing and scanning paper-and-pencil answer documents, (vii) delivering tests and scoring online versions of the EOCT, and

[1] The QCC has been undergoing transition to the Georgia Performance Standards (GPS). The GPS were introduced in spring 2005 in four content areas: ninth-grade literature and composition, American literature and composition, physical science, and biology. The GPS-based social studies EOCT of economics and U.S. history were administered for the first time in winter 2007. QCC is expected to be phased out by the end of the 2010-11 academic year.

(viii) preparing and distributing score reports, both on paper and online within a 5-day turnaround schedule.

The GHSGT is administered for the first time in the eleventh grade and covers five content areas: (i) English Language Arts; (ii) Mathematics; (iii) Science; (iv) Social Studies; and (v) Writing. The Writing assessment is administered each fall; the other four assessments are primarily administered during the spring assessment, with retest opportunities in the summer, fall, and winter. Pearson became the Georgia DOE's contractor for all GHSGT test development activities in January 2007. Prior to 2007, Pearson had been a subcontractor with responsibilities for printing test booklets, student answer documents, and other administration ancillary materials as well as for distributing and collecting test materials. The actual item writing, item content assignments, and answer key verification activities are subcontracted by Pearson to MP.

The contract between Pearson and the Georgia DOE states that Pearson will provide comprehensive program management for the GHSGT, oversee item development with the subcontractor, MP, provide psychometric services, including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the GHSGT program, design, print, and distribute all test materials and ancillary documents, including electronic and Braille test publishing, and prepare and distribute score reports.

The Examination Processing Review (EPR) summarized in this report was conducted exclusively as an off-site evaluation. An on-site visit to Pearson was originally scheduled to occur in December 2009, coinciding with Pearson's processing of the Georgia EOCT examinations from the winter administration. Unfortunately, severe weather in the Midwest during that time prevented CART staff from getting to Iowa City. The holidays and other resource-intensive, higher-priority activities (e.g., standard setting) further limited any attempts at re-scheduling the visit within the available processing window. It was therefore decided to conduct the entire PPSA Phase 2 evaluation off-site, with files downloaded/exchanged between Pearson staff and CART staff via a secure FTP server. CART staff received data and documentation for the processing of the EOCT examinations via a secure FTP server. This report summarizes CART's review of those data actually received by CART staff in response to the Data Request, an evaluation of empirical results, including re-analyses and audit checks on the data, and processing, quality assurance and quality control recommendations.

3.0 Examination Processing Review Results

The Examination Processing Review (EPR) was initiated in December 2009 with a formal Data Request sent to Pearson (see Appendix A). The Data Request outlines five types of results relative to PPS: (1) reconciliations of the scanned data; (2) post-administration test map verification and production of psychometric processing source files; (3) scoring table production; (4) individual score file production; and (5) census file scoring file production. Pearson's response to the Data Request (see Appendix B) was provided by mid-January 2010. Data files sent were primarily limited to the Economics EOCT test (forms [redacted] and [redacted] All files were provided to CART staff via a secure FTP server. In most cases, the files were samples of more complete data files. Some of the files (e.g., the incomplete score response data matrix and IRT calibration/equating results) were not available from the Winter 2009 administration[2].

Audit results, commentary, and recommendations are provided for each of the five areas noted above and further elaborated in the Data Requests (Appendix A).

3.1 Reconciliation of Scanned Data

As described in the CART Phase 1 audit report—see Luecht, R. M. & Ackerman, T. A., 2009, *Psychometric Procedures and Systems Audit: Data Management Review of Pearson for the Georgia End-of-Course Test (EOCT) and Georgia High School Graduation Test (GHSGT)*—the scanning operations at Pearson are state of the art with numerous quality control procedures in place to identify scanning discrepancies. It was expected that quality control reporting would be relatively straight-forward. That was not necessarily the case. Many of the summary (reconciliation) count reports requested from Pearson were apparently not readily available and much of requested data appeared to be difficult for them to compile. While it is acknowledged that the PPSA EPR and Data Request were outside of the scope of regular examination processing at Pearson, none of the

[2] Pearson uses a pre-equating model for the EOCT winter administration, where previously calibrated IRT item statistics are used to generate scores for the current sample of examinees. Calibration and equating data files are therefore not generated during the

requested information would seem to be unusual for a large-scale testing operation to provide.

Reconciliation reports and counts from scanning and initial processing, as outlined in the Data Request (Appendix A), were not specifically provided by Pearson. Instead, job-control language (JCL) coding (see Appendix C) was provided, along with a sample of 66 scan files from the Winter 2009 administration of the Georgia EOCT. There was no apparent, direct connection between production of the 66 files and the JCL coding. In fact, the JCL coding proved not to be at all useful for external review since most of the coding referred largely to unknown source files and intermediate processing operations (e.g., sorts using temporary variable fields). Notably, the JCL did appear to automatically generate emails when certain error conditions where encountered, but there were no clear reports or reconciliation counts generated from any of this code.

The 66 scan files were formatted as *denormalized* examinee records with raw item responses stored in separate column positions (i.e., forming a response *vector* for each examinee). This *denormalized* format of the scan files makes it somewhat difficult to reconcile any of the response data to specific items on test forms, without auxiliary knowledge of the fields (columns). That is, the same field (column position) can represent different items. We must then depend on a form reference identifier on each record and an item-to-test form map to resolve the unique items. Despite the apparent limitation, CART conducted various analyses of the sixty-six scan files. Table 1 presents a complete tabular report of the record-count results by scan file and across all of the files.

Table 1. Summary Counts for 66 Scan Files Provided by Pearson

Scan File	*Total Record Count*	*Complete Records*	*No-Name Records*	*Name-Only Records*	*Missing Name and Responses*
F0025900.SCAN001.TXT	1595	269	1224	100	2
F0026900.SCAN001.TXT	2039	882	1048	106	3
F0027900.SCAN001.TXT	2827	1358	1313	152	4
F0028900.SCAN001.TXT	2653	683	1855	112	3
F0029900.SCAN001.TXT	2953	1445	1326	178	4
F0030900.SCAN001.TXT	2947	1384	1423	137	3
F0031900.SCAN001.TXT	2619	494	1978	144	3
F0032900.SCAN001.TXT	2611	343	2134	129	5

F0033900.SCAN001.TXT	2596	322	2183	88	3
F0034900.SCAN001.TXT	2769	525	2084	157	3
F0035900.SCAN001.TXT	2823	1720	957	142	4
F0036900.SCAN001.TXT	2273	273	1888	109	3
F0037900.SCAN001.TXT	2579	1490	933	152	4
F0038900.SCAN001.TXT	2747	311	2286	146	4
F0039900.SCAN001.TXT	2567	461	1980	123	3
F0040900.SCAN001.TXT	3041	2194	694	150	3
F0041900.SCAN001.TXT	2846	682	1991	170	3
F0042900.SCAN001.TXT	1443	197	1177	67	2
F0043900.SCAN001.TXT	2648	388	2094	161	5
F0044900.SCAN001.TXT	2620	471	2006	138	5
F0045900.SCAN001.TXT	2664	493	2018	148	5
F0046900.SCAN001.TXT	791	208	530	52	1
F0047900.SCAN001.TXT	2527	356	2008	157	6
F0048900.SCAN001.TXT	2591	371	2113	104	3
F0049900.SCAN001.TXT	2686	597	1921	165	3
F0050900.SCAN001.TXT	2578	363	2054	158	3
F0051900.SCAN001.TXT	2724	714	1801	205	4
F0052900.SCAN001.TXT	2592	1539	896	150	7
F0053900.SCAN001.TXT	2738	1784	813	138	3
F0054900.SCAN001.TXT	2228	323	1826	76	3
F0055900.SCAN001.TXT	2546	969	1378	193	6
F0056900.SCAN001.TXT	2332	448	1727	154	3
F0057900.SCAN001.TXT	2034	544	1373	114	3
F0058900.SCAN001.TXT	1694	443	1128	121	2
F0059900.SCAN001.TXT	2652	1667	825	157	3
F0060900.SCAN001.TXT	2441	434	1886	115	6
F0061900.SCAN001.TXT	2640	2377	131	122	10
F0062900.SCAN001.TXT	2776	764	1889	120	3
F0063900.SCAN001.TXT	935	144	743	47	1
F0064900.SCAN001.TXT	2619	443	2057	116	3
F0065900.SCAN001.TXT	2170	1585	481	101	3
F0066900.SCAN001.TXT	3190	2809	253	125	3
F0067900.SCAN001.TXT	2573	790	1620	155	8
F0068900.SCAN001.TXT	2591	572	1873	143	3
F0069900.SCAN001.TXT	2669	2471	3	191	4
F0070900.SCAN001.TXT	2727	1153	1404	164	6
F0071900.SCAN001.TXT	2719	1995	569	151	4

F0072900.SCAN001.TXT	2760	539	2068	150	3
F0073900.SCAN001.TXT	2785	957	1588	236	4
F0074900.SCAN001.TXT	2646	1370	1144	129	3
F0075900.SCAN001.TXT	2620	833	1635	149	3
F0076900.SCAN001.TXT	2463	555	1745	160	3
F0077900.SCAN001.TXT	2604	983	1456	161	4
F0078900.SCAN001.TXT	2782	1033	1578	166	5
F0079900.SCAN001.TXT	2199	849	1238	107	5
F0080900.SCAN001.TXT	1691	267	1315	104	5
F0081900.SCAN001.TXT	2604	328	2135	138	3
F0082900.SCAN001.TXT	3	1	0	1	1
F0083000.SCAN001.TXT	83	59	8	15	1
F0084000.SCAN001.TXT	75	13	52	2	8
F0085000.SCAN001.TXT	39	6	21	10	2
F0086000.SCAN001.TXT	158	53	88	16	1
F0087000.SCAN001.TXT	135	126	0	8	1
F0088000.SCAN001.TXT	10	1	4	3	2
F0089000.SCAN001.TXT	1853	1734	45	71	3
F0090000.SCAN001.TXT	5	1	1	0	3
Sum	144138	51956	84015	7929	238
Percentage	100.0%	36.0%	58.3%	5.5%	0.2%

It is interesting to note that the totals at the bottom of Table 1 have no obvious reference to any known or expected quantities. For example, another score-reporting file generated by Pearson had 224,750 records, which actually exceeds the scan total count in these 66 files by approximately 80K examinee records[3]. From a simple count reconciliation perspective, it would be extremely useful to have external counts of all examinees to check the categories shown (complete records, partial records with no name or identifiers, partial records with incomplete or corrupted response data, and records with completely blank responses). We have no doubt that this type of detailed reconciliation information is available in Pearson's database systems. However, it: (a) salient summary reports do not appear to be routinely generated and (b), if generated, it

[3] An attempt was made to extract from the scan files only the records for the EOCT Economics forms 671 and 672 (apparent form field at column #27, width=3 characters). That analysis produced 24,323 examinee records, a count that did not even come close to the 29,961 examinee records subsequently extracted from the scoring records for those two test forms (see Section 3.2). The fact is that Pearson was unable to provide a detailed report listing scan record counts for every test form and for every item. We are therefore left to speculate as to what is the "correct" count.

is not clear who, if anybody, checks these data for discrepancies. For example, referring to Table 1, it appears that 58.3% of the scan file records had no name. There was a reference in the JCL to "virtual examinees", but this rather large percentage of essentially unidentified examinees is quite troubling from any reasonable data management perspective.

It seems useful to illustrate rather than merely describe what we mean by a *reconciliation report*. Appendix D includes a partial listing of a sample reconciliation report for an [unspecified] examination program with normalized data records (i.e., one examinee by form by item transaction per physical data record coming from either scanning or computer-based delivery). There were 506,100 examinee-by-item transactions for 3,857 items and 1,687 examinees. Including scrambled test forms, there were 72 unique test form identifiers. (Note: only a small sample of the actual results for actual 3,857 items are listed for illustrative purposes and in order to reduce the size of Appendix D.) Each item is cross referenced on all test forms on which the item appears. Counts of valid and discrepant examinee records are provided throughout the report, including at the item level. We are not suggesting that Pearson needs to mimic this report verbatim; but, this type of reconciliation and verified count summary information should be readily available, routinely checked, and certified for every examination processing cycle for the Georgia EOCT and GHSGT. The analyses and score reporting steps that follow are only as accurate as the source data provided. If only one examinee record is corrupted or otherwise not accounted for, or if one item is misidentified, a sound reconciliation report should pick up those discrepancies prior to initiating any psychometric analyses.

3.2 Post-administration Test Map Verification and Production of Psychometric Processing Source Files

Pearson provided five files relevant to the initial psychometric processing of the data (see Appendix B). All of the data were taken from the 80-item Economics EOCT, forms [redacted] and [redacted] . The TRIAN (item analysis) files were analyzed, and where possible, verified in several ways by CART.

First, using a source file provided by Pearson that contained 224,750 EOCT examinee records (file name: *Reporting file_01_12_10.txt*) and the answer keys contained in the Pearson TestMap item file (file name: *TestMap_EOCT_ECON_V02.xls*), the raw item responses in the former file were rescored and checked against the scored response

vectors in that same file. This check provided a confirmation of the answer keys. Second, the scored response records were used to identify an implicit key for each item (i.e., if the examinee gets a "1", the raw response must be a key).

Appendix E lists the summary reports for Economics forms [redacted] and [redacted] All of the scoring for form [redacted] (80 items, *N*=15,056 records extracted from the source file, *Reporting file_01_12_10.txt*) and for form [redacted] (80 items, *N*= 14,905 extracted records) matched with 100% accuracy for the key verification checks. Likewise, as shown in Appendix E, none of the items on either form had any "secondary" (implicit keys). We were unable to match these counts from the complete response data file against any of the item-level counts provided by Pearson for items appearing on these two test forms; that is, the file *TRIAN.FILE.EECO671.TXT_report.rtf* showed a count of 4,048 examinees including in the flagged items analysis on the paper form of [redacted] and *TRIAN.FILE.PECO671.TXT_report.rtf* reported a sample-based count of 2,574 examinees for the computer-based form.

3.3 Scoring Table Production

The test map file (*TestMap_EOCT_ECON_V02.xls*) contained Rasch item parameter estimates for two forms of EOCT Economics, [redacted] and [redacted] The estimated item difficulties for the operational items were used to independently compute proficiency scores for all possible raw score points on each test form. In turn, those proficiency scores (θ) were converted to scale scores, $y_{SS}=\beta 0 + \beta_1\theta$, using linear transformation constants provided by Pearson for the Economic examination (β_0=398.4717, β1=45.6204, lowest obtainable scale score or LOSS=200, and highest obtainable scale score or HOSS=650). Associated standard errors of estimate on the θ scale and transformed (rounded) to the scale score metric, $SE(y_{SS})=\beta_1 \times SE(\theta)$, were also independently computed. The results are provided in Appendix F for both forms [redacted] and [redacted]

The form [redacted] scoring table was compared to a scoring table provided by Pearson (file = eco_f1_RSSS_rounded.txt). That table from Pearson is reproduced in Appendix G. CART's results agree with the Pearson results to all reported decimal places, confirming the way in which the scoring tables were produced.

3.4 Individual Score File Production

The scored file *Reporting file_01_12_10.txt* was provided by Pearson without specific documentation—that is, without a data dictionary explaining the location of the data fields. No processing scripts or other scoring-relevant information were provided (see Data Request). Although test form identifiers and response vectors were apparent in that source file, it was not possible to verify if both raw scores and scale scores were even included.

Nonetheless, CART used the raw [rescored] responses to generate number-correct scores for the 68 operational items on two EOCT Economics forms ([redacted] and [redacted] The scoring tables recreated by CART (see Section 3.3) were then used to look up IRT proficiency scores (i.e., θ estimates), scale scores, and standard errors for both scores.

Appendix H provides the descriptive statistics separately summarizing by EOCT Economics test form the distributions of IRT θ score estimates, the asymptotic SE(θ) values, the scale scores, and the standard errors of the scale scores. Ideally, these distributional statistics could be matched to existing descriptive summaries from the Winter 2009 administration, verifying the actual score look-up process used by produce reported scores.

3.5 Final Census Score Table Production

CART did generate IRT proficiency scores (θ estimates[4]) and scale scores for 29,961 students who took the EOCT Economics forms ([redacted] and [redacted] see Section 3.4 and Appendix G). Since CART was not provided with any summary results from Pearson, there is unfortunately no basis for comparison to the numbers reported. It is not known if the data contained in *Reporting file_01_12_10.txt* is a census sample or an partial sample of the examinees who took the EOCT in Winter 2009.

4.0 Phase 2 Audit Conclusions and Recommendations

The response to the Data Request by Pearson appeared to be rather minimal. CART admittedly did not follow-up to get some of the specific missing data mentioned

[4] The Georgia EOCT and GHSGT are calibrated and scored using the Rasch model. Calibrations are performed using the WinSteps (Linacre, 2006) software. Unconditional maximum likelihood estimation is used for estimating both item difficulties and person proficiency scors.

through this audit summary. That is because it was anticipated that Pearson would have the data and reports readily available. The Data Request (see Appendix A) sent in December 2009 and was fairly explicit as to the nature of the data being requested. Pearson ultimately had over 30 days to respond to the Data Request and even that time line was considered flexible by joint agreement of CART staff, the Georgia DOE, and Pearson staff. There was also a follow-up telephone conference call between CART staff, Georgia DOE staff and Pearson staff, in December, to discuss content of the Data Request, with an implied commitment by Pearson to provide the data, if available. Ultimately, CART agreed to a lower priority treatment of the Data Request so that Pearson could devote all relevant resources to standard setting activities in January 2010.

Although Pearson staff members were seemingly very cooperative, ultimately, the data and reports provided by Pearson for the Phase 2 audit seemed to be woefully insufficient as delineated throughout this document. While it may have made sense, on the surface, for CART to follow-up to request additional information, the intent of the Phase 2 audit was to verify what is currently done operationally and routinely in processing the Georgia EOCT and GHSGT. We therefore decided to proceed with the Phase 2 audit using the limited data and document deficiencies for possible consideration in Phase 3.

The good news was that all of CARTs checks, including the re-analysis of any available data, verified Pearson's results with 100% accuracy. The more unfortunate news is that some seemingly straight-forward checks and reconciliations were not possible because comparative counts were not available and/or raw data were not provided. The apparent lack of routine count reconciliation data throughout the examination processing cycle is somewhat troubling. Responsible individuals and companies carefully reconcile their checking accounts on a regular basis to ensure agreement between the funds they believe that they have on hand and what a bank shows as their balance. The same should be true for examination data. Georgia needs the assurance that every examinee response to every item is fully accounted for and represented in the results reported by Pearson. Reconciliation reports and verified count summary information should be readily available at any point within a processing cycle for an entire examination program, a discipline (e.g., EOCT Economics), or other established query criteria. Equally important, those reconciliation results should be standard operating procedure and routinely checked and certified for every step in the examination processing cycle for the EOCT and GHSGT.

In addition to better reconciliation data, it would also be useful for Pearson to define (and document) one or more target analysis samples and to use those same samples for all subsequent analyses. This approach might avoid reporting different sample-based statistics from time-specific, query-based extractions from a dynamic data repository. The queries used to generate the analysis samples and locked images of the data need to be stored to ensure that results can be exactly reproduced.

For example, Pearson provided two sample item analysis results, *TRIAN.FILE.EECO671.TXT_report.rtf* and *TRIAN.FILE.PECO671.TXT_report.rtf*, showing distractor analysis results and classical item statistics for three items, each computed from different samples (one from the computer-based test administration and one from the paper-and-pencil test). The statistical results are different, which makes perfect sense, given that they are computed using different examinee samples. The question is, however, whether Pearson would be able to reproduce both sets of results from source files now or at some point in the future. Clearly, it would not be sufficient to state that those statistics were estimated using paper-and-pencil and computer-based versions of the EOCT Economics test form [redacted] since the total count of 6,622 examinees used in those item analyses is also different than the 15,056 form [redacted] examinees identified by CART from the file *Reporting file_01_12_10.txt*. Those types of discrepancies need to be documented and explained[5]. For example, the 6,622 examinees may be a subset of the 15,056 examinees that happened to receive that pretest block of items.

[5] Other states do engage a third party to carry out a confirmatory review, possibly with re-analysis of the primary examination processing steps.

Appendix A

CART PPSA Phase 2 Data Request

09 December 2009

1. *Scanned data reconciliation* (comprehensive counts report prepared by Pearson). Purpose: to verify that all scanned data reconciles to expectation insofar as numbers of answer sheet processed and examinee response records created. The reconciliation report should include counts with respect to the following:
    a. Valid (active) test forms with counts of items (scored, unscored, total)
    b. Numbers of examinee answer sheets and response records created and verified for each test form
    c. Number of examinees per item (possible flags for low-count or zero-count items)
    d. Item counts by test forms with subcounts for operational items, equating items, pretest items, and unscored items (e.g., survey items)
    e. Item counts per examinee (total, scored, pretest, omits), matched to expected counts for that form (report minimum, maximum, mean, std. deviation of item counts across examinee records)
2. *Post-administration test map verification and documentation and production of psychometric analysis source files*. Purpose: to ensure the correct production of scored responses used in calibration, scaling, and final scoring. Scope: one paper-and-pencil test form and one CBT form with data for a random sample of 1,000 examinees—see "d" and "e", below
    a. Final IA and key validation results (Pearson IA)
    b. Identification of any miskeyed items (or approval of form). Note: if any forms have miskeyed items, please focus on those.
    c. Item answer key source table used in raw scoring
    d. Raw response file (normalized: person + item id + raw response): samples of 1,000 examinees (per form)
    e. Incomplete data matrix (IDM) scored response file used for calibration; i.e., provide actual Pearson-generated IDM files—but include IDs so that we can match to the raw responses
    f. Summary report(s) from actual Pearson generation of the Fall 2009 IDM files

3. *Fall 2009 scoring table production results*. Purpose: to verify the extraction of banked item statistics, generation of score tables, and computation (look-ups) of scale scores for a random sample of examinees. Scope: one paper-and-pencil test form and one CBT form with data for a random sample of 1,000 examinees—please provide a different sample than from #2, above.
    a. Equated IRT item statistics (equated WinSteps Rasch item difficulties) for the two test forms
    b. Pearson-generated score tables and scale-score calculations (raw-score to theta and theta to scale score)
    c. Samples of 1000 randomly chosen examinees taking each type of form (different than #2, above)
4. *Individual score file production*. Purpose: to verify the production sequence of the final, individual score file(s) uploaded to GaDOE.
    a. Roster-level score table production scripts
    b. Summary checks on results (e.g., statistical summaries)
    c. Sign-offs and lock-down
    d. Uploading to GaDOE
    e. Verification that uploaded file is final, locked-version
    f. Census score file and summary reports production. Purpose: to verify the production sequence of the final scored file(s) uploaded to GaDOE.
5. Final census score table production. Purpose: to verify the production sequence of the final, score file(s) and summary reports uploaded to GaDOE.
    a. Production scripts
    b. Summary checks on results (e.g., statistical summaries)
    c. Sign-offs and lock-down
    d. Uploading to GaDOE
    e. Verification that uploaded file is final, locked-version

Appendix B

Pearson Response to the Data Request

CART PPSA Phase 2 Data Request

09 December 2009

2. *Post-administration test map verification and documentation and production of psychometric analysis source files*. Purpose: to ensure the correct production of scored responses used in calibration, scaling, and final scoring. Scope: one paper-and-pencil test form and one CBT form with data for a random sample of 1,000 examinees—see "d" and "e", below
    a. Final IA and key validation results (Pearson IA)—key checks

        TRIAN files for economics, core form 1, for both paper and online forms, TRIAN.FILE.PECO671.TXT (paper form of economics) and TRIAN.FILE.EECO671.TXT (online form of economics) and key check report files (TRIAN.FILE.EECO671.TXT_report.rtf and TRIAN.FILE.PECO671.TXT_report.rtf) are provided[6]. The key check report files include items that are flagged using a set of statistical criteria. These items then were sent to content for key review. This process is run for all subjects for both testing modes.

        Final IA is usually not carried out with TRIAN. Psychometrics does compute final item-level statistics for all operational items for a given administration and present these statistics as part of the annual technical report.

    b. Identification of any miskeyed items (or approval of form). Note: if any forms have miskeyed items, please focus on those.

        Key check report files (containing items that are flagged using a set of statistical criteria) are sent to content for key checks. Content will then sign off on the accuracy of the keys. For this past winter administration, no items were considered miskeyed items through this process.

    c. Item answer key source table used in raw scoring

[6] Because no subject was specified in the file request document, we randomly picked Economics as the subject we provided files for. If files for other subjects are preferred, please let us know and we can provide.

Customer test map is provided for economics. File name: TestMap_EOCT_ECON_V02.xls. Form [redacted] contains relevant information for the core form 1.

d. Raw response file (normalized: person + item id + raw response): samples of 1,000 examinees (per form)

From IT

e. Incomplete data matrix (IDM) scored response file used for calibration; i.e., provide actual Pearson-generated IDM files—but include IDs so that we can match to the raw responses

EOCT adopts a pre-equating model and no IDM is used for operational equating. Therefore, no IDM was produced for the winter 2009 administration, and hence not provided.

f. Summary report(s) from actual Pearson generation of the Fall 2009 IDM files

No IDM was generated for the winter 2009 (fall 2009) administration (see e above).

3. *Fall 2009 scoring table production results*. Purpose: to verify the extraction of banked item statistics, generation of score tables, and computation (look-ups) of scale scores for a random sample of examinees. Scope: one paper-and-pencil test form and one CBT form with data for a random sample of 1,000 examinees—please provide a different sample than from #2, above.

a. Equated IRT item statistics (equated WinSteps Rasch item difficulties) for the two test forms

Online and paper are the same form offered on both testing modes. The test map provided under 2(c) contains equated IRT item parameters (variable name: Rasch).

b. Pearson-generated score tables and scale-score calculations (raw-score to theta and theta to scale score)

eco_f1_RSSS_rounded.txt contains the raw to theta to scale score table for core form 1. This scoring table was applied to both paper and online forms.

The item parameters used to generate the table are listed on the test map (variable name: Rasch). Here is the linear transformation used:

Content Area	LOSS	HOSS	Slope	Intercept	Scale Score Transformation
Economics	200	650	45.6204	398.4717	$45.6204 * \theta + 398.4717$

Here is the layout file for the RSSS table:

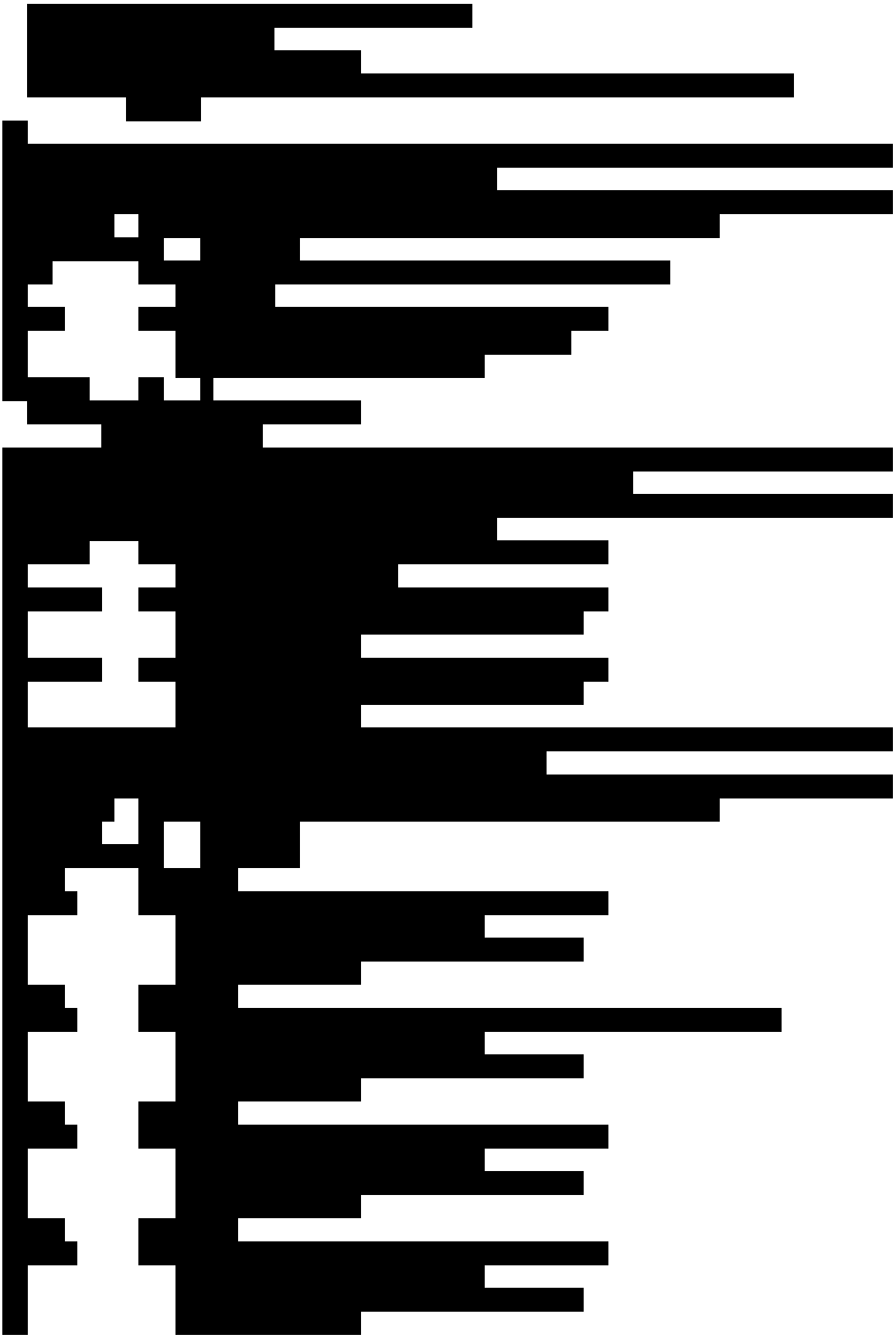
Variable Description	Columns	Variable Length	SAS Format	Example
Subject	1-11	11	$11.	
Form Number	20-24	5	$5.	
Braille form or not	27	1	$1.	0 – regular form 1 – Braille form
Raw Score	31-32	2	2.	
Scale Score	35-37	3	3.	
Scale Score CSEM	41-43	3	3.	
Theta	50-57	8	8.5	
Theta CSEM	70-77	8	8.5	

c. Samples of 1000 randomly chosen examinees taking each type of form (different than #2, above)

From IT

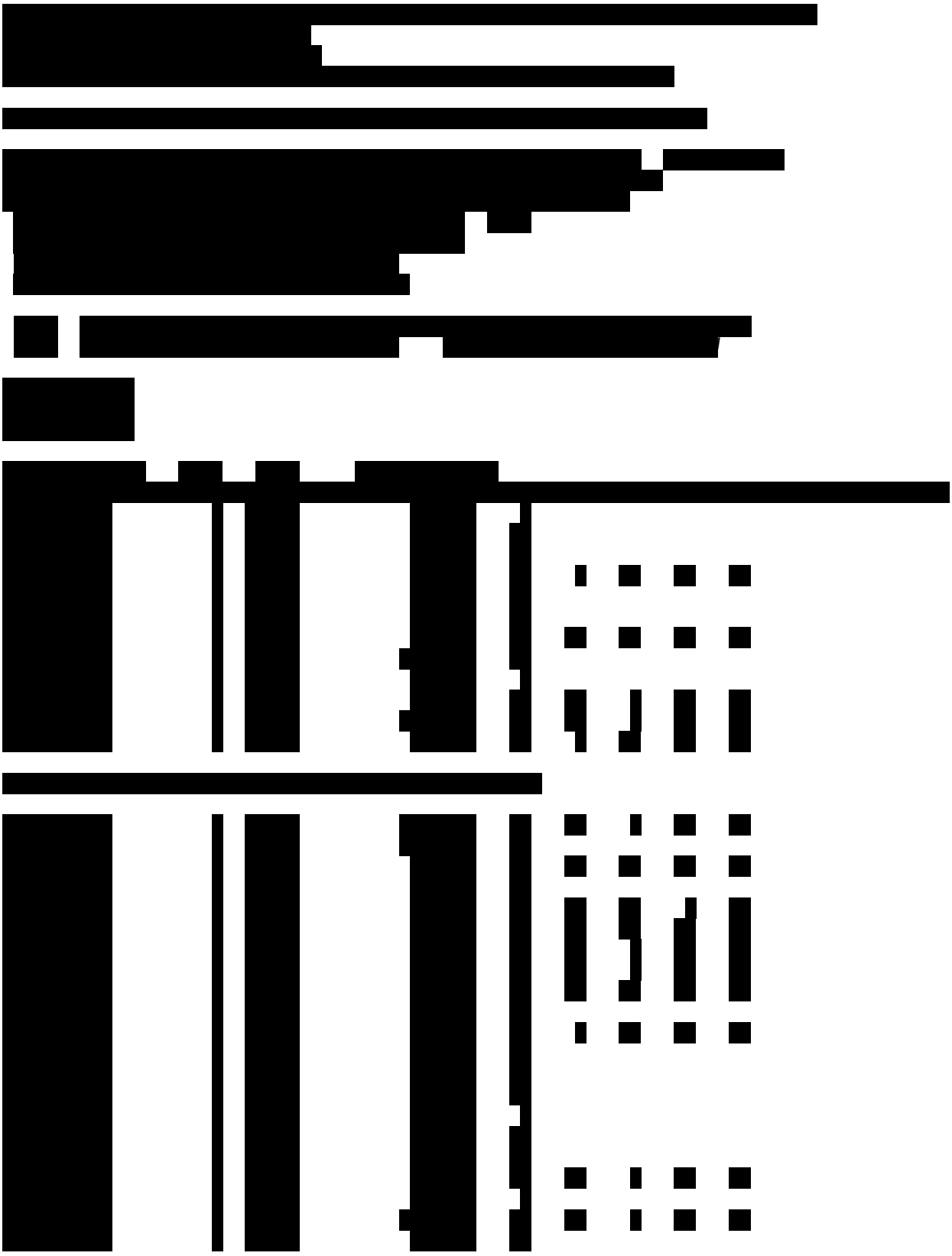
Appendix C

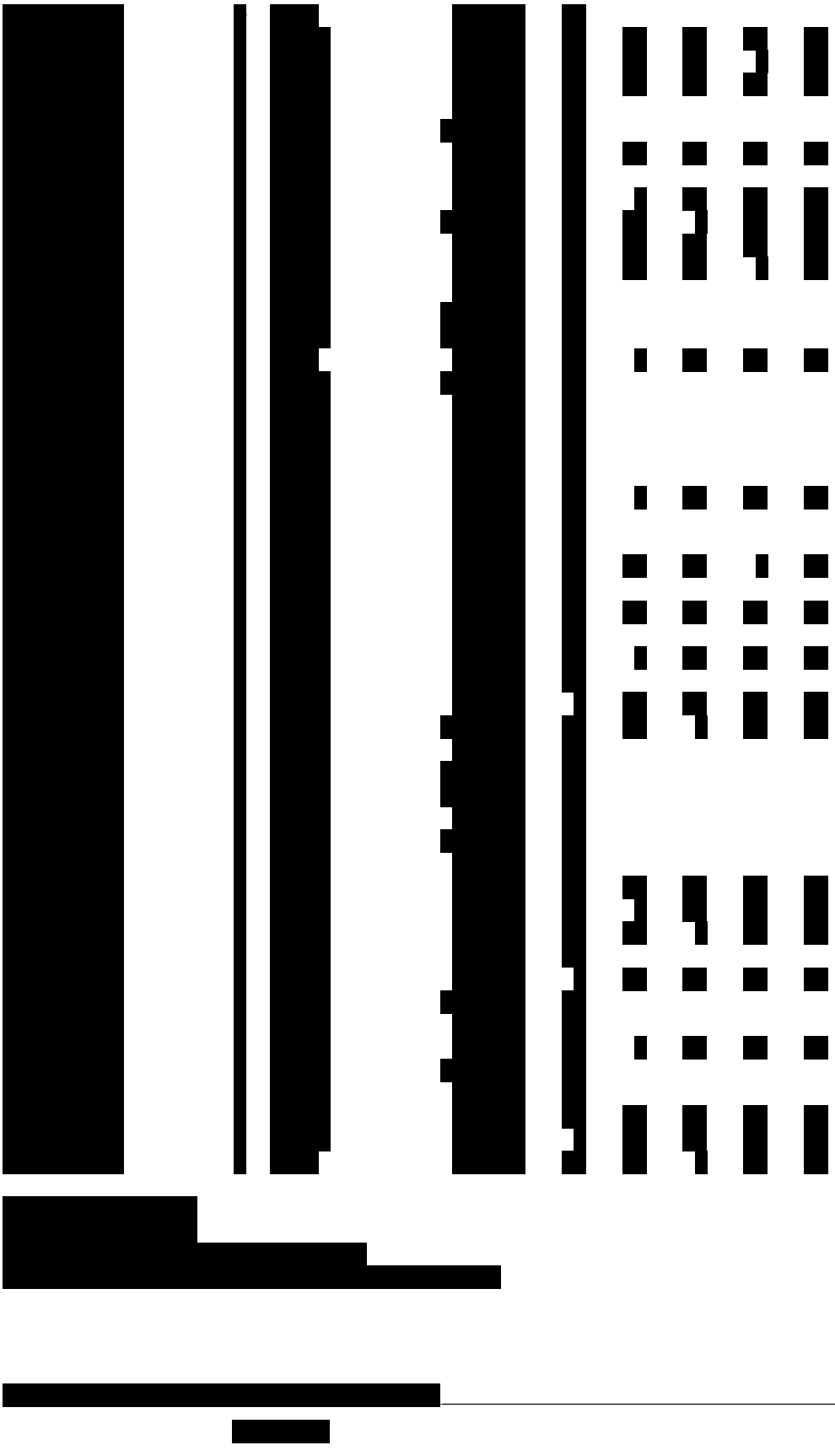
Sample JCL from Pearson

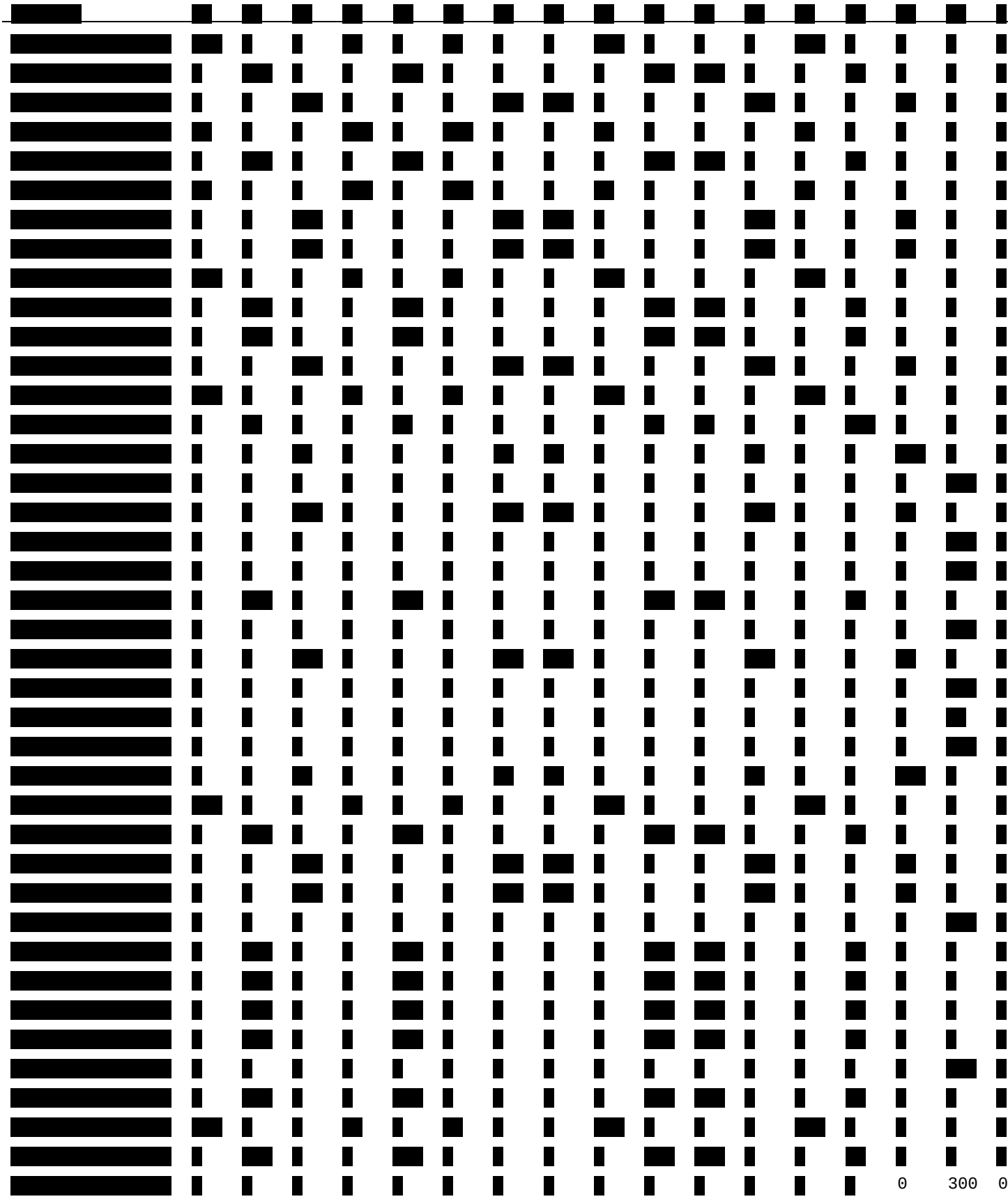


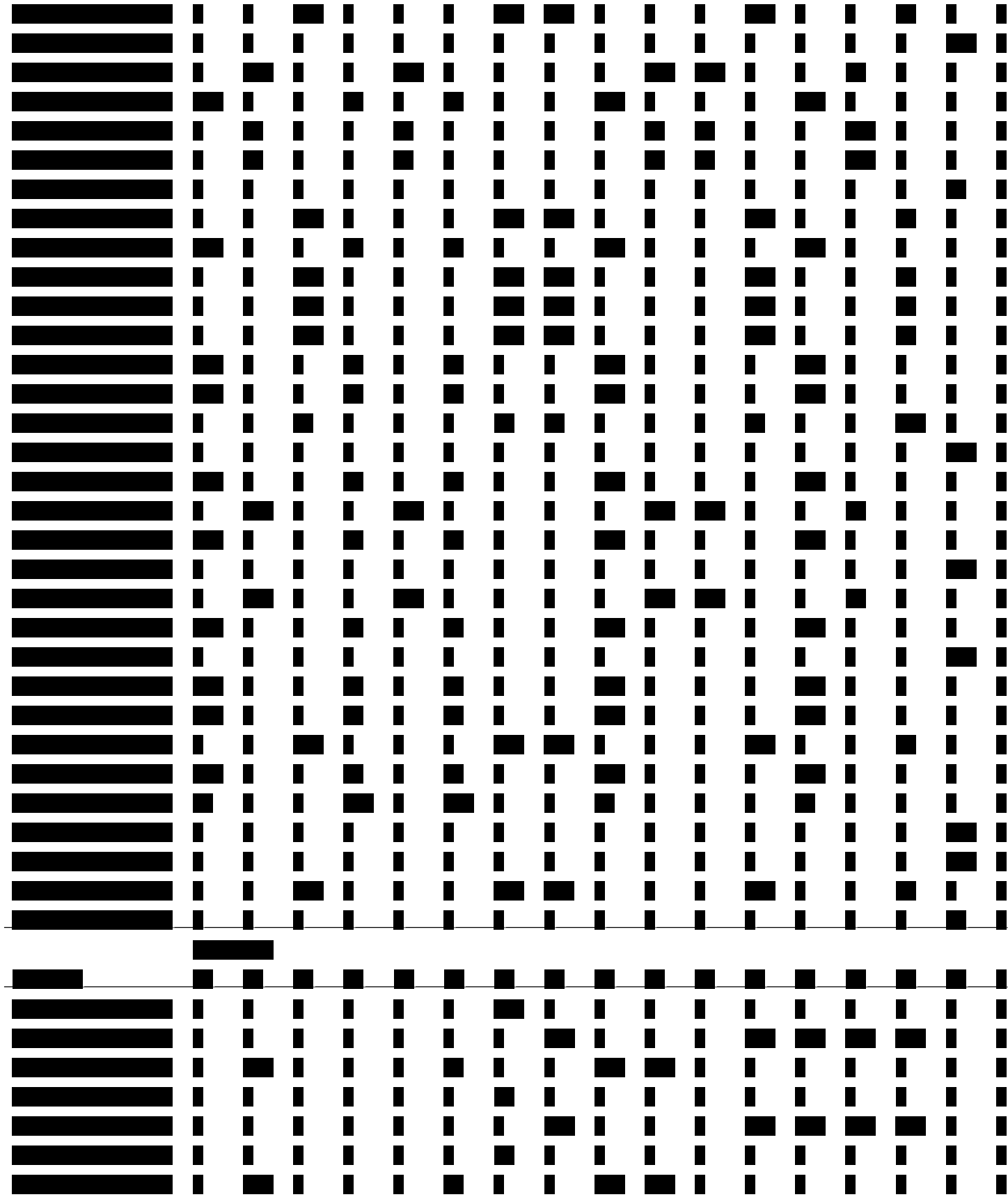
Appendix D

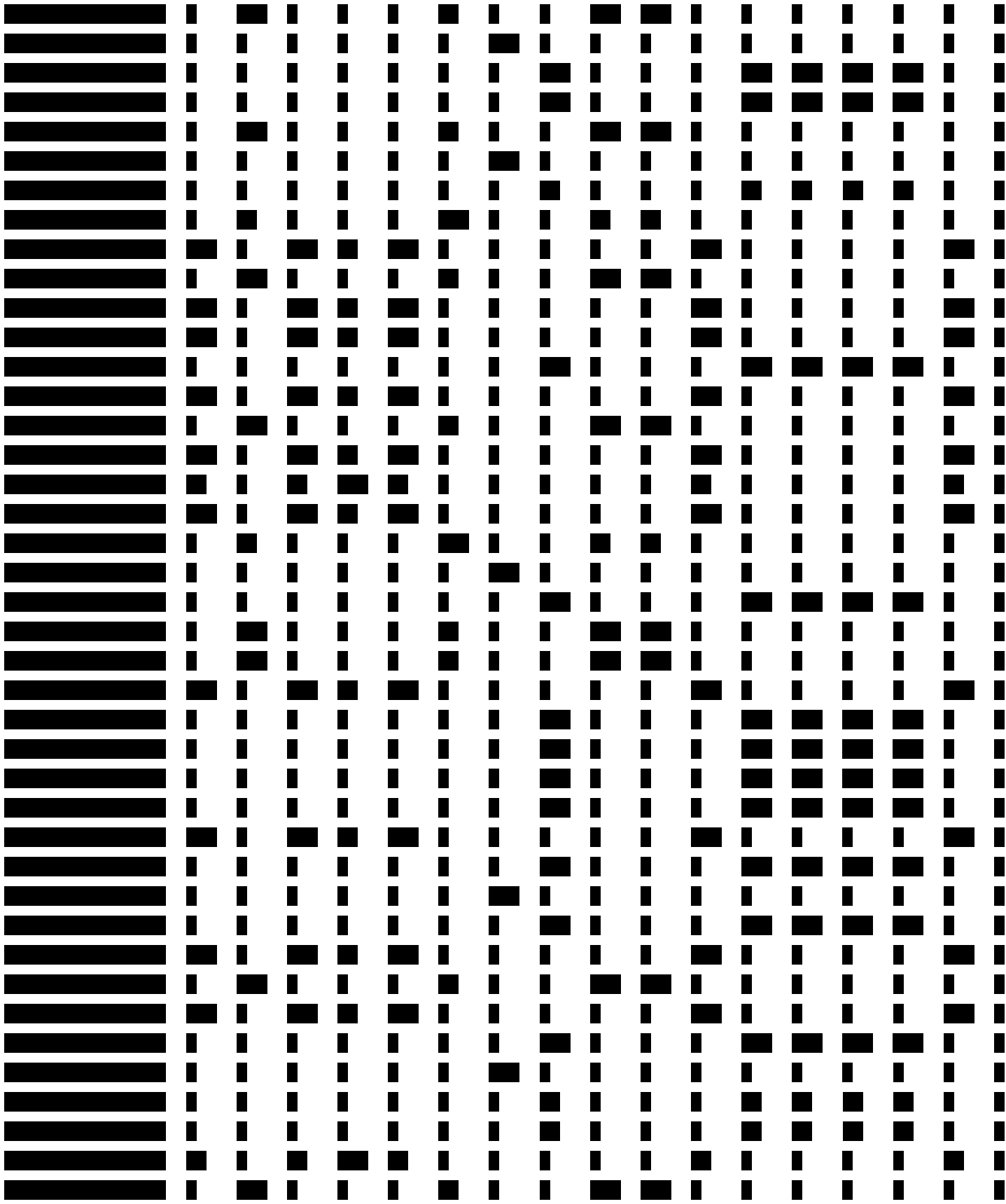
Sample CART Reconciliation Report (Item List is Partial)

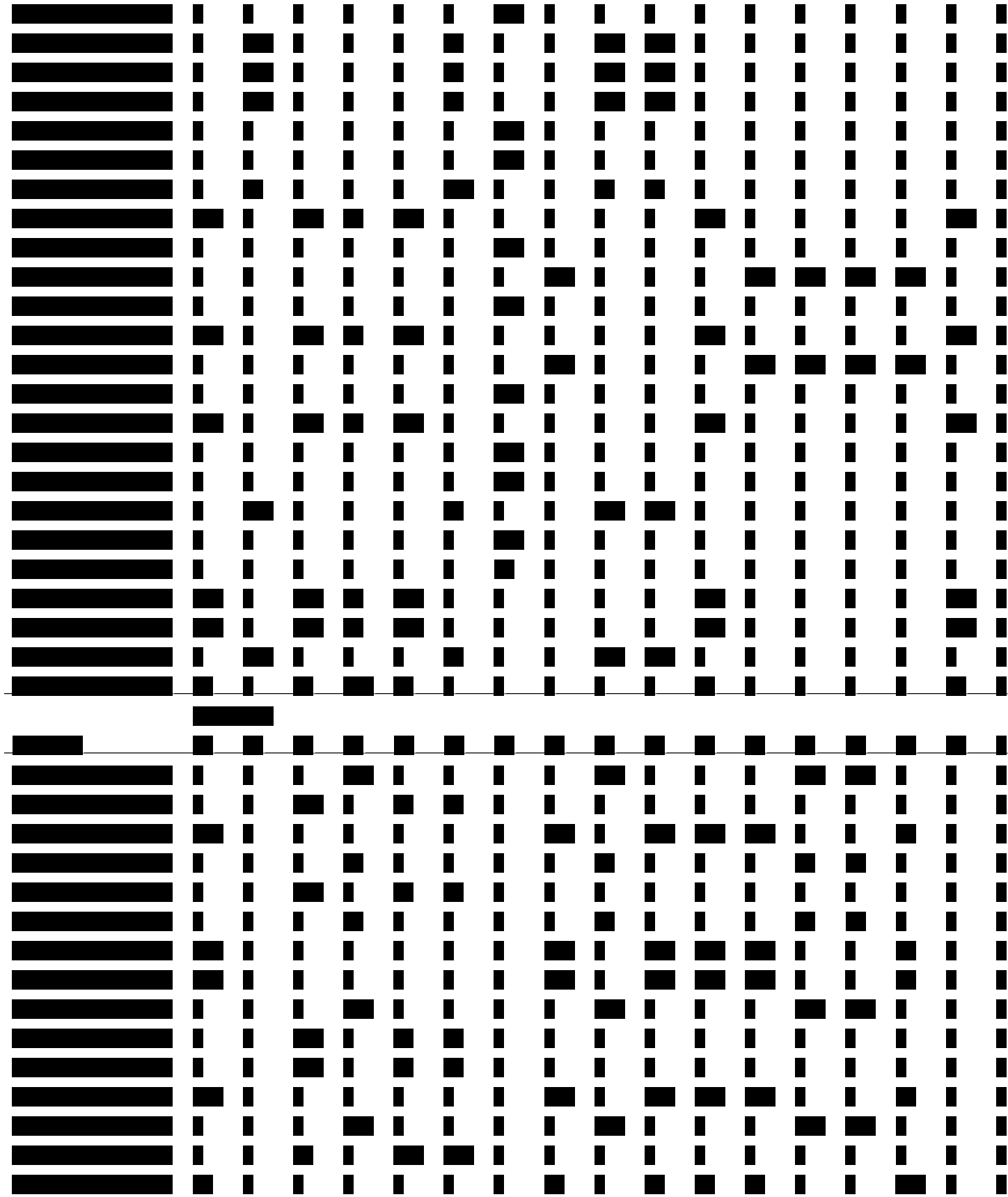


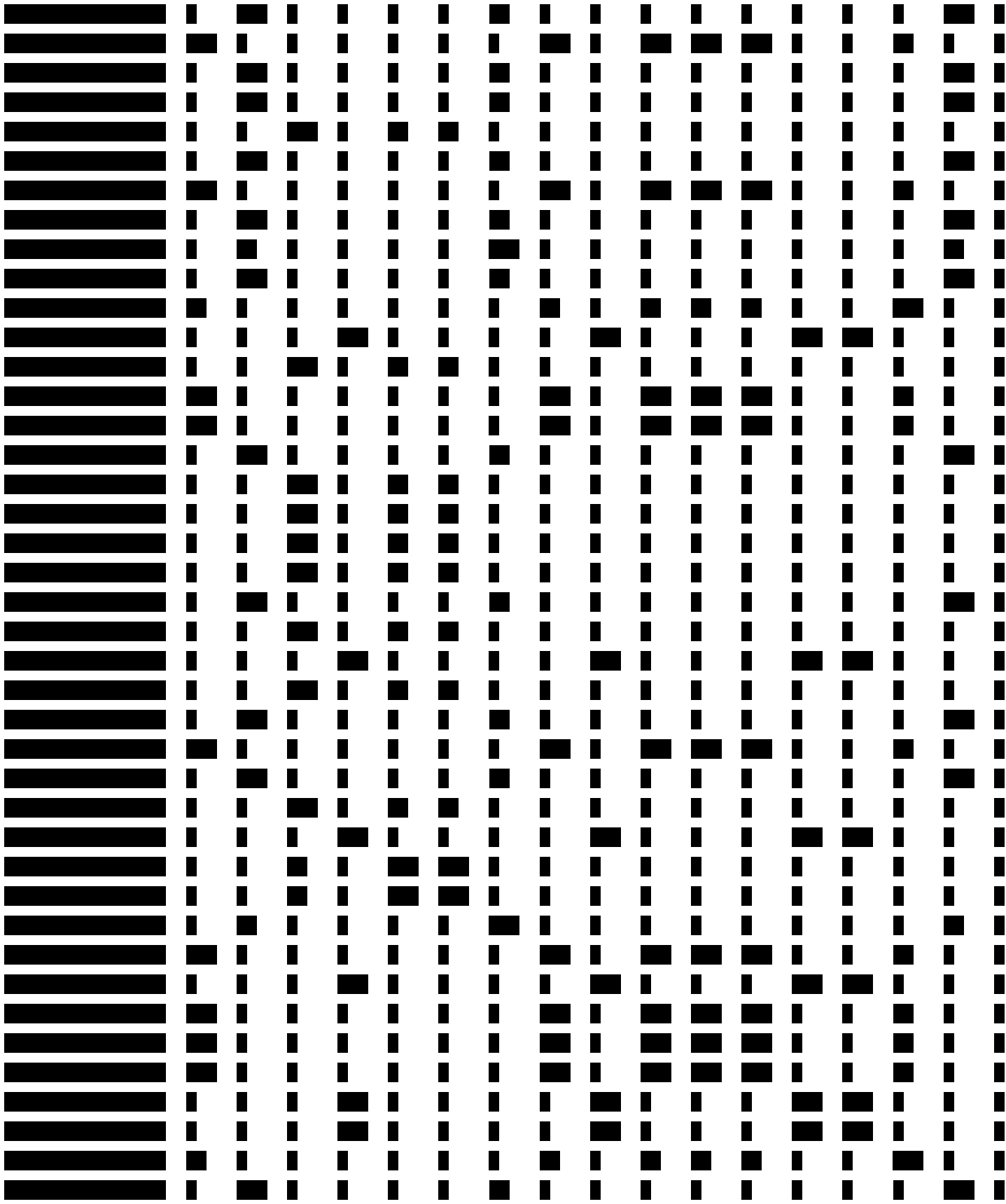


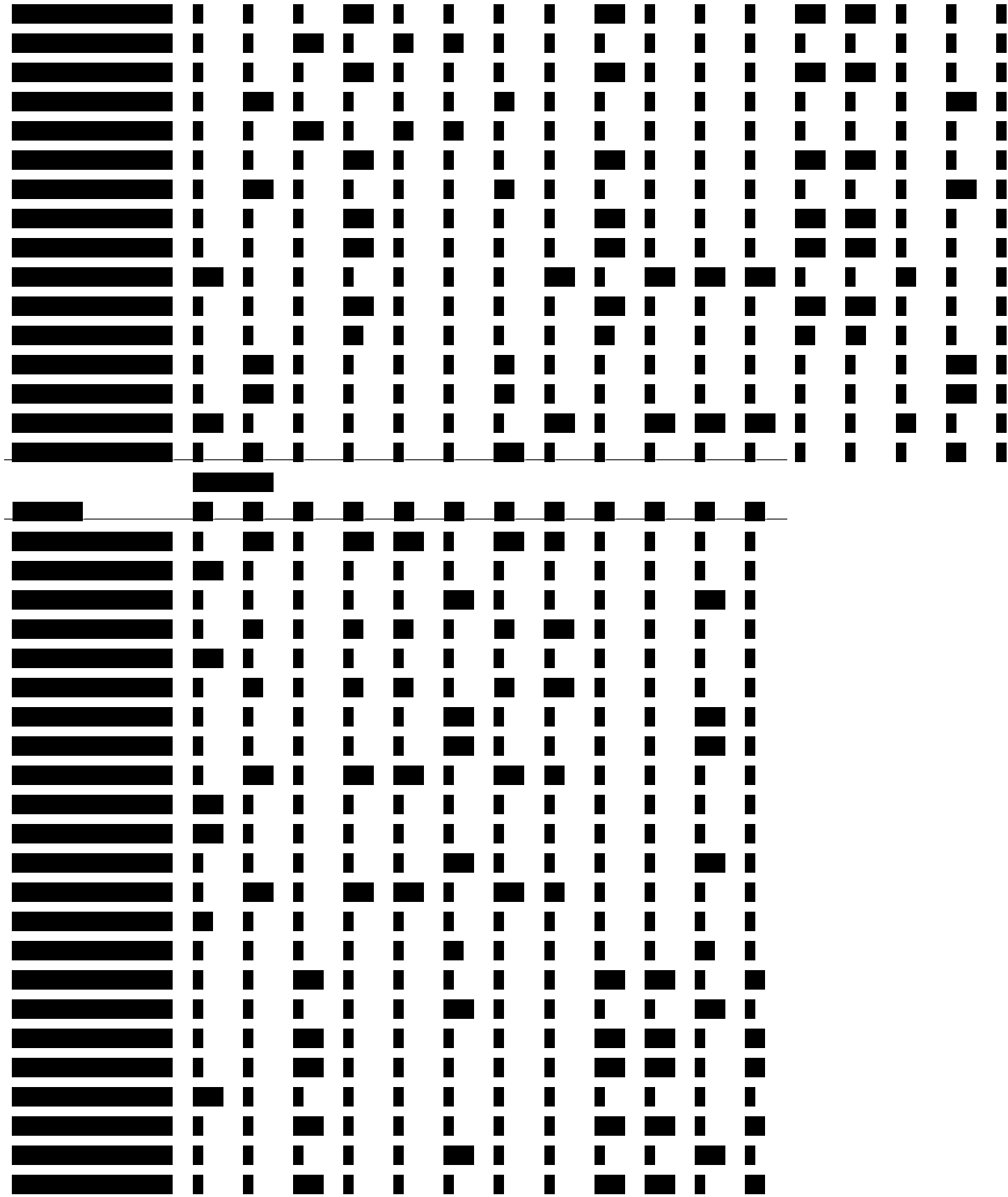

0
300
0

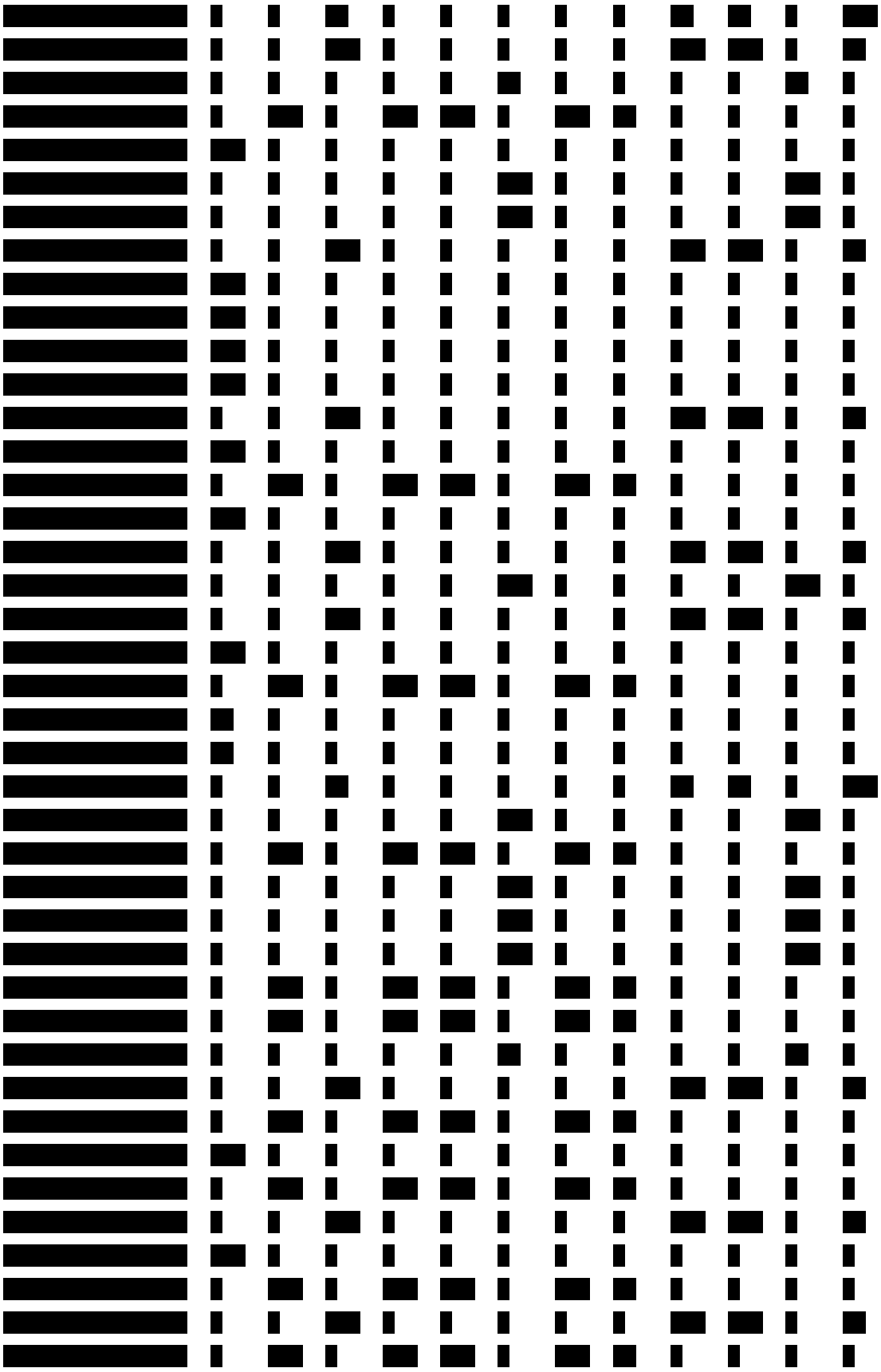


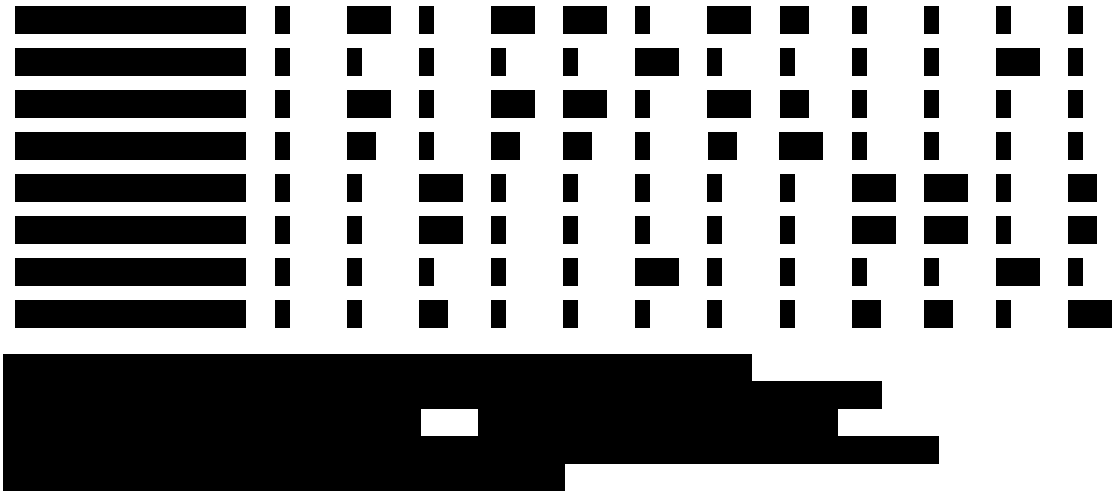








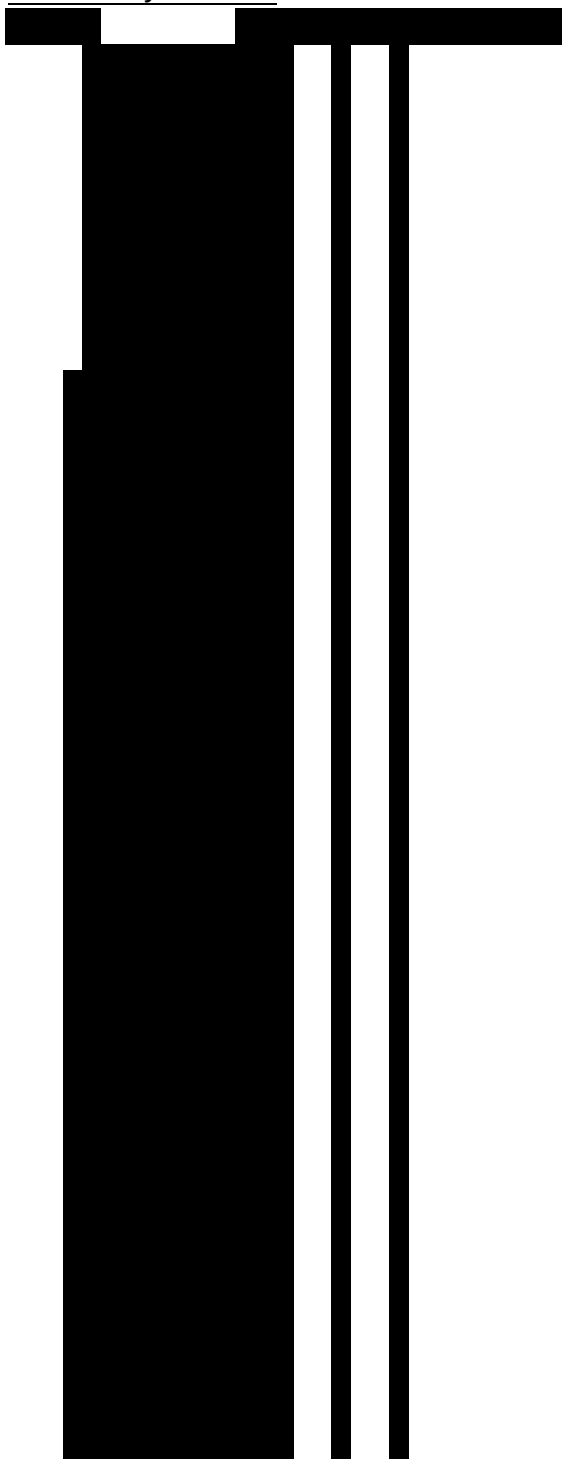


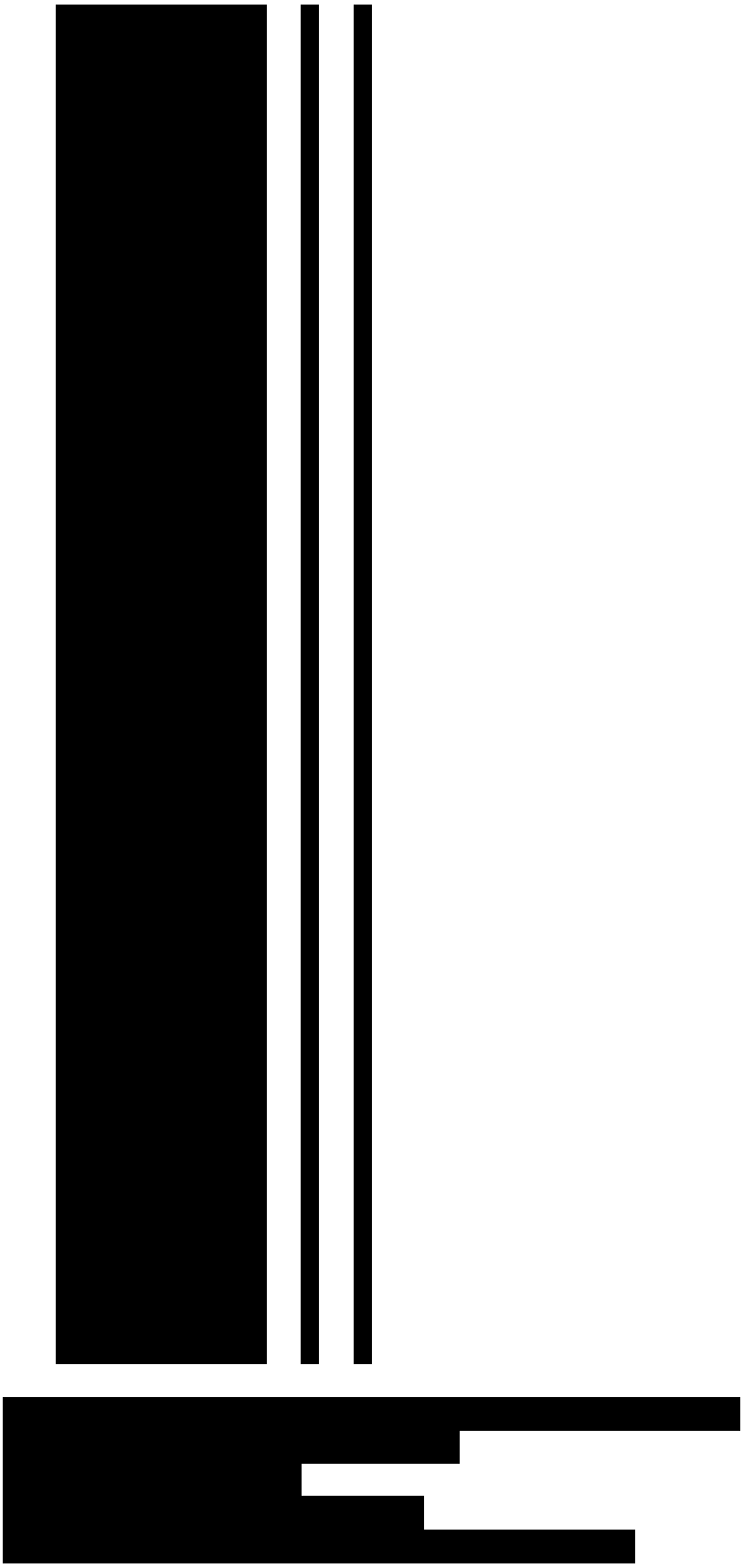


Appendix E

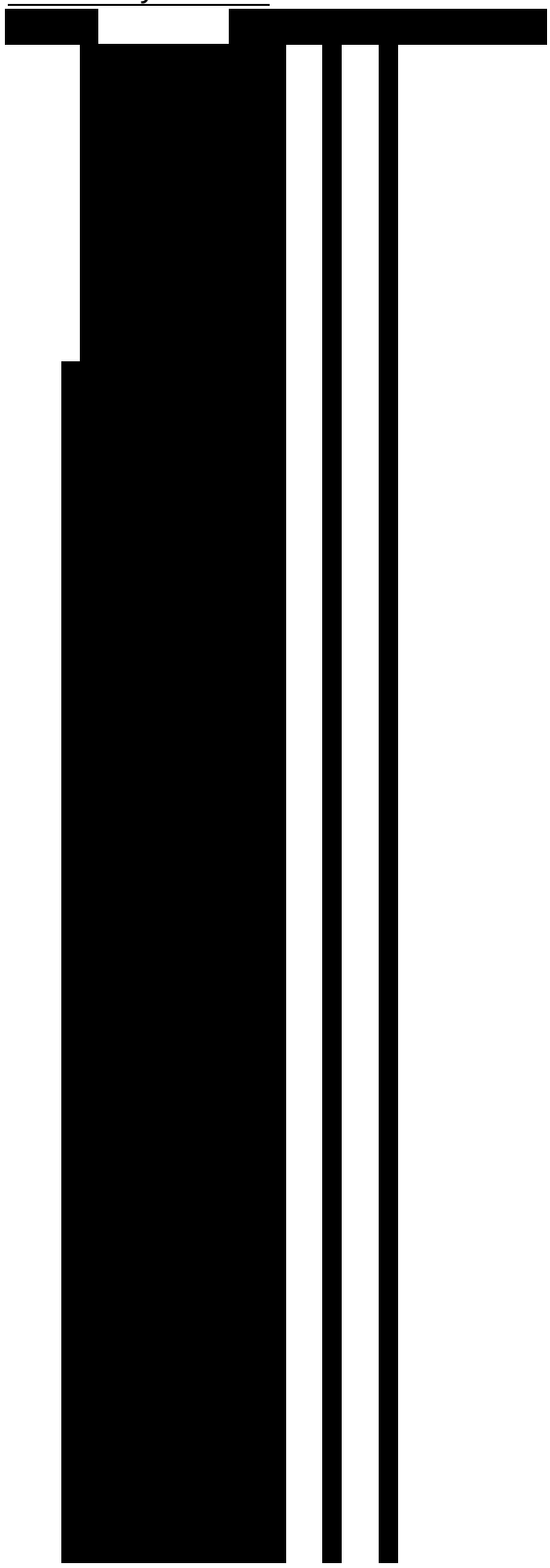
Summary of Examinee Counts for Economics Forms [redacted] and [redacted]

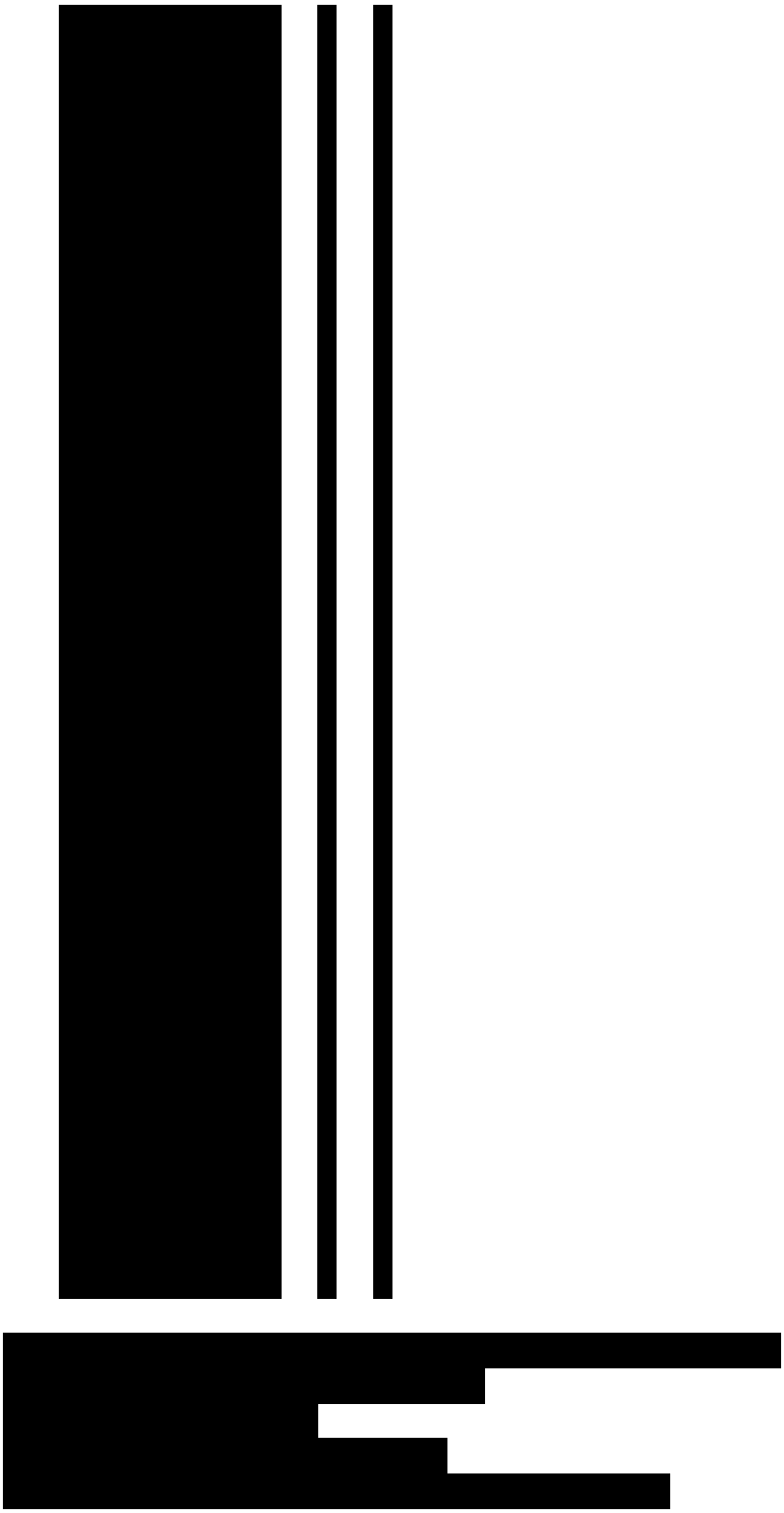
```
Item Answer Key File: [redacted]_KeyStatus.csv
NI= 80
Item Key Check
```





Item Answer Key File: [redacted]_KeyStatus.csv
NI= 80
Item Key Check





(Note: _KeyStatus.csv files extracted intact from *TestMap_EOCT_ECON_V02.xls*)

Appendix F

Scoring Tables for Forms [redacted] and [redacted] (Produced by CART)

	Economics Form [redacted]					Economics Form [redacted]			
Raw Score	Pct. Score	Theta, θ	SE(θ)	Scale Score	SE(y_{SS})	Theta, θ	SE(θ)	Scale Score	SE(y_{SS})
0	0.0	-5.6790	1.8320	200	84	-5.6790	1.9770	200	90
1	1.5	-4.4591	1.0110	200	46	-4.3050	1.0110	202	46
2	2.9	-3.7431	0.7230	228	33	-3.5901	0.7220	235	33
3	4.4	-3.3147	0.5970	247	27	-3.1627	0.5960	254	27
4	5.9	-3.0038	0.5230	261	24	-2.8528	0.5220	268	24
5	7.4	-2.7574	0.4730	273	22	-2.6073	0.4720	280	22
6	8.8	-2.5515	0.4360	282	20	-2.4024	0.4350	289	20
7	10.3	-2.3737	0.4080	290	19	-2.2254	0.4070	297	19
8	11.8	-2.2162	0.3860	297	18	-2.0688	0.3850	304	18
9	13.2	-2.0742	0.3680	304	17	-1.9277	0.3670	311	17
10	14.7	-1.9445	0.3530	310	16	-1.7988	0.3520	316	16
11	16.2	-1.8245	0.3400	315	16	-1.6796	0.3390	322	15
12	17.7	-1.7126	0.3290	320	15	-1.5684	0.3280	327	15
13	19.1	-1.6073	0.3200	325	15	-1.4639	0.3190	332	15
14	20.6	-1.5077	0.3120	330	14	-1.3650	0.3100	336	14
15	22.1	-1.4129	0.3040	334	14	-1.2709	0.3030	340	14
16	23.5	-1.3222	0.2980	338	14	-1.1809	0.2970	345	14
17	25.0	-1.2351	0.2920	342	13	-1.0945	0.2910	349	13
18	26.5	-1.1511	0.2870	346	13	-1.0112	0.2860	352	13
19	27.9	-1.0698	0.2830	350	13	-0.9305	0.2820	356	13
20	29.4	-0.9909	0.2790	353	13	-0.8522	0.2780	360	13
21	30.9	-0.9141	0.2750	357	13	-0.7760	0.2740	363	13
22	32.4	-0.8391	0.2720	360	12	-0.7016	0.2710	366	12
23	33.8	-0.7657	0.2700	364	12	-0.6288	0.2680	370	12
24	35.3	-0.6937	0.2670	367	12	-0.5574	0.2660	373	12
25	36.8	-0.6230	0.2650	370	12	-0.4872	0.2640	376	12
26	38.2	-0.5534	0.2630	373	12	-0.4181	0.2620	379	12
27	39.7	-0.4846	0.2610	376	12	-0.3499	0.2600	383	12
28	41.2	-0.4167	0.2600	379	12	-0.2825	0.2590	386	12
29	42.7	-0.3494	0.2590	383	12	-0.2157	0.2580	389	12
30	44.1	-0.2827	0.2580	386	12	-0.1494	0.2570	392	12
31	45.6	-0.2164	0.2570	389	12	-0.0835	0.2560	395	12

32	47.1	-0.1504	0.2570	392	12	-0.0180	0.2560	398	12
33	48.5	-0.0846	0.2560	395	12	0.0474	0.2560	401	12
34	50.0	-0.0190	0.2560	398	12	0.1127	0.2550	404	12
35	51.5	0.0467	0.2560	401	12	0.1780	0.2560	407	12
36	52.9	0.1124	0.2570	404	12	0.2434	0.2560	410	12
37	54.4	0.1784	0.2570	407	12	0.3091	0.2570	413	12
38	55.9	0.2446	0.2580	410	12	0.3751	0.2570	416	12
39	57.4	0.3113	0.2590	413	12	0.4415	0.2580	419	12
40	58.8	0.3785	0.2600	416	12	0.5085	0.2590	422	12
41	60.3	0.4463	0.2610	419	12	0.5761	0.2610	425	12
42	61.8	0.5149	0.2630	422	12	0.6446	0.2620	428	12
43	63.2	0.5844	0.2650	425	12	0.7140	0.2640	431	12
44	64.7	0.6550	0.2670	428	12	0.7845	0.2670	434	12
45	66.2	0.7268	0.2690	432	12	0.8562	0.2690	438	12
46	67.7	0.8000	0.2720	435	12	0.9294	0.2720	441	12
47	69.1	0.8748	0.2750	438	13	1.0042	0.2750	444	13
48	70.6	0.9514	0.2790	442	13	1.0808	0.2790	448	13
49	72.1	1.0301	0.2830	445	13	1.1596	0.2830	451	13
50	73.5	1.1112	0.2870	449	13	1.2408	0.2870	455	13
51	75.0	1.1949	0.2920	453	13	1.3247	0.2920	459	13
52	76.5	1.2817	0.2980	457	14	1.4117	0.2980	463	14
53	77.9	1.3721	0.3040	461	14	1.5024	0.3040	467	14
54	79.4	1.4665	0.3110	465	14	1.5972	0.3120	471	14
55	80.9	1.5658	0.3190	470	15	1.6968	0.3200	476	15
56	82.4	1.6707	0.3290	475	15	1.8021	0.3290	481	15
57	83.8	1.7822	0.3400	480	15	1.9142	0.3400	486	16
58	85.3	1.9017	0.3520	485	16	2.0343	0.3530	491	16
59	86.8	2.0310	0.3670	491	17	2.1642	0.3680	497	17
60	88.2	2.1724	0.3850	498	18	2.3064	0.3860	504	18
61	89.7	2.3293	0.4080	505	19	2.4641	0.4090	511	19
62	91.2	2.5066	0.4350	513	20	2.6424	0.4370	519	20
63	92.7	2.7118	0.4720	522	22	2.8486	0.4730	528	22
64	94.1	2.9576	0.5220	533	24	3.0956	0.5230	540	24
65	95.6	3.2677	0.5960	548	27	3.4070	0.5970	554	27
66	97.1	3.6953	0.7230	567	33	3.8362	0.7230	573	33
67	98.5	4.4105	1.0110	600	46	4.5530	1.0120	606	46
68	100.0	5.6297	1.8320	650	84	5.6297	1.7070	650	78

Appendix G

Scoring Tables for Form [redacted]

(Pearson File: *eco_f1_RSSS_rounded.txt*)

(Header rows added)

Subject	Form	Brl	Raw Scor	Scale Scor	Scale CSEM	Theta	Theta CSSEM
ECONOMICS	[redacted]	0	0	200	84	-5.67900	1.83100
ECONOMICS		0	1	200	46	-4.45910	1.01100
ECONOMICS		0	2	228	33	-3.74310	0.72300
ECONOMICS		0	3	247	27	-3.31470	0.59700
ECONOMICS		0	4	261	24	-3.00380	0.52200
ECONOMICS		0	5	273	22	-2.75740	0.47200
ECONOMICS		0	6	282	20	-2.55150	0.43600
ECONOMICS		0	7	290	19	-2.37370	0.40800
ECONOMICS		0	8	297	18	-2.21620	0.38600
ECONOMICS		0	9	304	17	-2.07420	0.36800
ECONOMICS		0	10	310	16	-1.94450	0.35200
ECONOMICS		0	11	315	16	-1.82450	0.34000
ECONOMICS		0	12	320	15	-1.71260	0.32900
ECONOMICS		0	13	325	15	-1.60730	0.31900
ECONOMICS		0	14	330	14	-1.50770	0.31100
ECONOMICS		0	15	334	14	-1.41290	0.30400
ECONOMICS		0	16	338	14	-1.32220	0.29800
ECONOMICS		0	17	342	13	-1.23510	0.29200
ECONOMICS		0	18	346	13	-1.15110	0.28700
ECONOMICS		0	19	350	13	-1.06980	0.28200
ECONOMICS		0	20	353	13	-0.99090	0.27900
ECONOMICS		0	21	357	13	-0.91410	0.27500
ECONOMICS		0	22	360	12	-0.83910	0.27200
ECONOMICS		0	23	364	12	-0.76570	0.26900
ECONOMICS		0	24	367	12	-0.69370	0.26700
ECONOMICS		0	25	370	12	-0.62300	0.26400
ECONOMICS		0	26	373	12	-0.55340	0.26300
ECONOMICS		0	27	376	12	-0.48460	0.26100
ECONOMICS		0	28	379	12	-0.41670	0.26000
ECONOMICS		0	29	383	12	-0.34940	0.25800
ECONOMICS		0	30	386	12	-0.28270	0.25700
ECONOMICS		0	31	389	12	-0.21640	0.25700
ECONOMICS		0	32	392	12	-0.15040	0.25600
ECONOMICS		0	33	395	12	-0.08460	0.25600
ECONOMICS		0	34	400	12	-0.01900	0.25600
ECONOMICS		0	35	401	12	0.04670	0.25600
ECONOMICS		0	36	404	12	0.11240	0.25600
ECONOMICS		0	37	407	12	0.17840	0.25700
ECONOMICS		0	38	410	12	0.24460	0.25700
ECONOMICS		0	39	413	12	0.31130	0.25800
ECONOMICS		0	40	416	12	0.37850	0.25900

ECONOMICS	[redacted]	0	41	419	12	0.44630	0.26100
ECONOMICS		0	42	422	12	0.51490	0.26200
ECONOMICS		0	43	425	12	0.58440	0.26400
ECONOMICS		0	44	428	12	0.65500	0.26600
ECONOMICS		0	45	432	12	0.72680	0.26900
ECONOMICS		0	46	435	12	0.80000	0.27200
ECONOMICS		0	47	438	13	0.87480	0.27500
ECONOMICS		0	48	442	13	0.95140	0.27800
ECONOMICS		0	49	445	13	1.03010	0.28200
ECONOMICS		0	50	450	13	1.11120	0.28600
ECONOMICS		0	51	453	13	1.19490	0.29100
ECONOMICS		0	52	457	14	1.28170	0.29700
ECONOMICS		0	53	461	14	1.37210	0.30300
ECONOMICS		0	54	465	14	1.46650	0.31100
ECONOMICS		0	55	470	15	1.56580	0.31900
ECONOMICS		0	56	475	15	1.67070	0.32800
ECONOMICS		0	57	480	15	1.78220	0.33900
ECONOMICS		0	58	485	16	1.90170	0.35200
ECONOMICS		0	59	491	17	2.03100	0.36700
ECONOMICS		0	60	498	18	2.17240	0.38500
ECONOMICS		0	61	505	19	2.32930	0.40700
ECONOMICS		0	62	513	20	2.50660	0.43500
ECONOMICS		0	63	522	22	2.71180	0.47200
ECONOMICS		0	64	533	24	2.95760	0.52200
ECONOMICS		0	65	548	27	3.26770	0.59600
ECONOMICS		0	66	567	33	3.69530	0.72200
ECONOMICS		0	67	600	46	4.41050	1.01000
ECONOMICS		0	68	650	84	5.62970	1.83100

Appendix H

Descriptive Summary of Converted Scores (from Scoring Look-up Tables) for Forms [redacted] and [redacted]

	Theta, θ	SE(θ)	Scale Score, y_{SS}	SE(y_{SS})
		Form [redacted]		
N of cases	15056	15056	15056	15056
Minimum	-5.6790	0.2560	200.0000	12.0000
Maximum	5.6297	1.8320	650.0000	84.0000
Mean	0.4645	0.2908	419.7679	13.3571
Standard Dev	1.0186	0.0778	46.3329	3.5604
		Form [redacted]		
N of cases	14905	14905	14905	14905
Minimum	-5.6790	0.2560	200.0000	12.0000
Maximum	5.6297	1.8320	650.0000	84.0000
Mean	0.3024	0.2809	412.3807	12.9114
Standard Dev	0.9125	0.0591	41.5342	2.7032

Psychometric Procedures and Systems Audit:
Phase 3 Real-Time Processing and Implementation Audit
for the Georgia End-of-Course Test (EOCT)
and Georgia High School Graduation Test (GHSGT)

Richard M. Luecht, PhD

Terry Ackerman, PhD

Center for Assessment and Research Technology (CART)

Greensboro, North Carolina

July, 2010

1.0 Introduction

The CART Psychometric Procedures and Systems Audit (PPSA) is a three-phase, technical evaluation of the end-to-end flow of item and examinee data. The purpose of the audit is to evaluate many aspects of the assessment processing system, from test development through scoring and reporting. The PPSA summarized here was carried out on behalf of the Georgia Department of Education, in cooperation with the Educational Measurement Group of Pearson in Iowa City, Iowa. All phases of the PPSA are limited to Pearson operations as they apply to the Georgia End-of-Course Tests (EOCT) and the Georgia High School Graduation Test (GHSGT).

Phase 1 of the PPSA is referred to as the Data Management Review (DMR). The DMR phase of the PPSA carried out for the EOCT and GHSGT was a formal, technical evaluation of Pearson's documented data systems, data management structures and processing procedures for items, test forms, and examination results (*see Psychometric Procedures and Systems Audit: Data Management Review of Pearson for the Georgia End-of-Course Test (EOCT) and Georgia High School Graduation Test (GHSGT)*, authored by Drs. Richard Luecht and Terry Ackerman of the Center for Assessment and Research Technology (CART), November 2009). The DMR further evaluated the integrity of the Georgia data as it flows throughout the Pearson systems, including test-form and item data management, test administration and processing of raw results, back-end examinee results processing psychometric analysis procedures and scoring, software and data management systems improvements, and quality control/assurance steps. The DMR concluded with formal recommendations regarding potential weaknesses in those systems and possible solutions or improvements that could help ensure the integrity of the data and processing steps. Phase 2 of the audit is called the Examination Processing Review (EPR). The EPR is an operational, technical evaluation of the results generated by various software programs and manual procedures used in end-to-end processing of an examination. Phase 2 was completed in spring 2010 and summarized in the technical report, *Psychometric Procedures and Systems Audit: Examination Processing Review for the Georgia End-of-Course Test (EOCT) and Georgia High School Graduation Test (GHSGT)* authored by R. M. Luecht and T. A. Ackerman (March 2010).

This report summarizes the Phase 3 Real-Time Processing and Implementation Audit. (RTPIA). The primary activity during the RTPIA phase of the PPSA was an unscheduled site visit to Pearson's main processing location to observe and evaluate procedures during actual examination processing. The RTPIA was similar in intent to the Phase 2 EPR, but focused more on evaluating observable activities during actual

processing. The RTPIA also included a review and discussion with key staff as to the progress made by Pearson in response to the CART Phase 1 and 2 audit memos and recommendations

2.0 Scope of the Phase 3 RTPIA

The PPSA covers and evaluation of Pearson's data management systems and processing procedures related to the Georgia End-of-Course Tests (EOCT) and the Georgia High School Graduation Test (GHSGT). Pearson's contractual obligations for test development, administration, and processing are summarized below.

The Georgia EOCT is administered in grades nine through twelve for eight state-mandated core subjects: (i) Mathematics I (Algebra, Geometry, and Statistics); (ii) Mathematics II (Geometry, Algebra II, and Statistics); (iii) U.S. History; (iv) Economics (including Business and Free Enterprise); (v) Biology; (vi) Physical Science; (vii) Ninth Grade Literature and Composition; and (viii) American Literature and Composition. In addition, legacy Algebra I and Geometry test forms are administered to accommodate students who entered high school under the previously authorized Georgia Quality Core Curriculum (QCC). Any student taking an EOCT course, regardless of grade level, is required to take the corresponding EOCT upon completion of that course. EOCT scores are averaged in as 15% of each student's final course grade. New EOCT tests are usually constructed for winter and spring administrations. Recycled tests are used for the summer administration and benchmark mid-month administrations. The EOCT can be administered via paper-and-pencil assessments or in an online format. Paper-and-pencil assessments are only available during the winter, spring or summer administrations. The EOCT items are developed by Measured Progress (MP) in collaboration with the Georgia DOE staff and high school educators. The test forms are designed and developed by Pearson staff, who also take full contractual responsibility for the following activities: (i) providing comprehensive program management, (ii) overseeing item development by MP, (iii) providing psychometric services including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the EOCT program, (iv) creating customized administration procedures for receipt control, data editing, and scoring processes, (v) designing, printing, and distributing all test materials and ancillary documents, including electronic and Braille test versions, (vi) processing and scanning paper-and-pencil answer documents, (vii) delivering tests and scoring online versions of the

EOCT, and (viii) preparing and distributing score reports, both on paper and online within a 5-day turnaround schedule.

The GHSGT is administered for the first time in the eleventh grade and covers five content areas: (i) English Language Arts; (ii) Mathematics; (iii) Science; (iv) Social Studies; and (v) Writing. The Writing assessment is administered each fall; the other four assessments are primarily administered during the spring assessment, with retest opportunities in the summer, fall, and winter. Pearson became the Georgia DOE's contractor for all GHSGT test development activities in January 2007. Prior to 2007, Pearson had been a subcontractor with responsibilities for printing test booklets, student answer documents, and other administration ancillary materials as well as for distributing and collecting test materials. The actual item writing, item content assignments, and answer key verification activities are subcontracted by Pearson to MP.

The contract between Pearson and the Georgia DOE states that Pearson will provide comprehensive program management for the GHSGT, oversee item development with the subcontractor, MP, provide psychometric services, including item analysis, scoring table generation, data review, standard setting, and other psychometric activities related to the GHSGT program, design, print, and distribute all test materials and ancillary documents, including electronic and Braille test publishing, and prepare and distribute score reports.

The Phase 3 Real-Time Processing and Implementation Audit (RTPIA) had the benefit of two prior phases: the DMR and EPR. The DMR and EPR reports provided a number of CART recommendations regarding possible improvements to Pearson's data management /processing systems and human/manual procedures as they pertain to ensuring the integrity of the Georgia EOCT and GHSGT data and scores. Pearson was advised prior to the RTPIA that those recommendations would be in primary focus during the visit.

The primary activity carried out for the RTPIA phase was an unannounced site visit to Pearson's Iowa City facilities in May 2010 by a senior member of the CART PPSA team. As alluded to above, in addition to observing the live examination processing activities at Pearson the RTPIA included to the degree possible verification of all corrective actions/implemented improvements by Pearson that addressed problems or recommendations raised during the first two phases of the PPSA.

3.0 Real-Time Processing and Implementation Audit Results

Pearson was informed[1] on 11-May-2010 that the RTPIA would take place on 17-May-2010. Dr. Terry Ackerman, a senior member of the CART PPSA team, visited the Pearson facilities in Iowa City, Iowa. Pearson prepared an agenda for the visit (see Appendix A). A narrative summary of the RTPIA visit follows.

The visit started on the morning of 17-May-2010 with an introductory meeting between Terry Ackerman and key Pearson staff, including Shannon Still, Michelle Klingeman and Chris Skapyak. The purpose of this introductory meeting was to outline the various areas at Pearson that would be visited and observed throughout the day. The onsite RTPIA was divided up into four components: (1) a review of the overall Georgia Quality Action Plan (GQAP); (2) an observational tour of the processing areas at Pearson currently involved in processing the Georgia examinations; (3) a discussion about the software development and enhancements implemented to address issues and recommendations discussed in previous PPSA reports; and (4) a review and evaluation of of the various elements and systems involved in the program management for Georgia.

The Georgia Quality Action Plan (see Appendix B) was provided to CART and the Georgia DOE in response to CART's Phase 1 DMR report. The components of the GQAP were discussed with Chris Skapyak and Dadi Narasimba. Both individuals elaborated how the GQAP had been implemented, including mandatory meetings and internal quality reviews and checks by various team members and other groups involved in processing the Georgia examinations at Pearson. Based upon those discussions, Pearson appears to have implemented all of the steps, improvements and deadlines identified in their GQAP. However, no evidence was provided to CART by Pearson that specifically evaluated the effectiveness of those procedural changes and quality control (QC) checks. For example, Pearson could have provided summary results as identified in the *Metrics and Tracking* section of the GQAP, but did not do so. We will return to this point in Section 4.0 of this report.

The tour of the processing facilities was conducted by Chris Skapyak and Michelle Klingeman. The various teams and systems involved in the Georgia program

[1] Although technically "unannounced", by agreement of all concerned parties, staff members at Pearson informed of the visit six days in advance to ensure that the company would have the appropriate staff on hand and available to accommodate the visit, without slowing down or otherwise interrupting ongoing spring administration examination processing for the Georgia programs.

appeared to be very well coordinated. There were with numerous tangible checks and backups procedures in place to help ensure the integrity of the physical examination data. As noted in CART's Phase 1 DMR report, many of Pearson's systems for materials handling and associated processing are state-of-the-art from a manufacturing perspective. The tour provided the opportunity to observe these procedures first hard. For example, once examination materials are received at the Pearson facility, processing of the answer sheets is carefully tracked, reconciled and monitored to ensure proper handling through scanning and encoding of the examinee response data. The answer sheets are logged in and then moved to storage rooms that controlled for humidity and temperature,, with numerous physical checks made to ensure that all answer sheets can be scanned. Likewise, the scanning equipment was state-of-the art and appeared to be well maintained and monitored. No problems were apparent in getting the response data into the system[2].

It made little sense to visit computer operators and analysts working on the data for several reasons. First, it would have been impossible to coordinate specific processing activities or analyses with the limited time frame of the RTPIA visit. Second, it would probably have been disruptive to have Pearson's operators and analysts take time to actually show their processing steps—possibly needing to walk the CART auditor through rather tedious steps and explain what was being done. Instead, a review and discussion of the software-related processing steps was undertaken.

The third component of the RTPIA visit involved a software review and overview meeting with Pearson's software, QC and program management staff (Chris Skapyak, Brad Russell, Dadi Narasimba, Shannon Still and Michelle Klingman). Program management/oversight activities and implementation of most of the QC steps for the Georgia examination programs are handled by the Assessment and Information Quality (AIQ) group. The AIQ group also has direct responsibility to improve quality and response time during processing, including troubleshooting anomalous conditions. Changes authorized by the implemented by Pearson's Software and Technology (S&T) group. Based on discussions with members of the AIQ and S&T, it appears that Pearson has added new procedures to improve the fidelity of the processing steps.

[2] As recommended in the Phase 1 DMR report, there is no attempt to reconcile printed test booklets and answer sheets originally shipped to each school district with the counts of materials returned to Pearson. This presents a minor security loop hole in the system that could be rectified by having the districts conduct their own exact reconciliation of all materials and at least enter counts of materials sent back to Pearson to ensure a match between the physical counts of answer sheets and what the districts contend was sent.

Many of these improvements were detailed during the meeting (also see Appendix B). The meeting also included a discussion of Pearson's Asset Management System (AMS), a comprehensive planning and QC system for monitoring workflows and system improvements. Based on the discussion, it appears that Pearson now uses highly scripted, systematic procedures that are constantly being reviewed and revised to make the end-to-end data processing more efficient and accurate.

The final component of the RTPIA involved a discussion of the results of the previous PPSA phases and information that was shared with CART. This discussion included Terry Ackerman, Chris Skapyak and Shannon Still. It was pointed out that software control code, data and reports that were provided to CART by Pearson—especially for the Phase 2 EPR—were not accompanied with any usable empirical results or information, and thus largely precluded any effective comparisons. For example, verifications of examinee- and item-count reconciliations could not be performed because Pearson provided no reconciliation reports. Likewise software code provided by Pearson was of no help in the EPR phase of the audit because no explanation/purpose for the code was provided. Pearson acknowledged the deficits in the information provided. The lack of available empirical information as to reconciliations and quality control checks remains a theme that we again address in our section 4.0 recommendations.

In summary, the visit to Pearson for the third phase of the CART audit was insightful and reassuring for the most part. The RTPIA visit clearly demonstrated a serious and professional commitment by Pearson to continually improve their systems and provide better service to the state of Georgia. As noted above, there are numerous cross-checks and other QC procedures to insure the accuracy of the scoring and reporting of test results. Many of those procedures were put in place in 2010 in direct response to the CART Phase 2 recommendations regarding automated processing and QC (see Appendix B). In that regard, the tension at Pearson between *proactive* QC and systemic change and *reactive* fixes to new issues and problems regarding the Georgia examination programs seems to be finally balanced more in favor the proactive side of systems design. However, there is still room for improvement (see section 4.0 of this report).

Section 4.0 Summary Comments and Recommendations

As noted in section 3.0, the RTPIA proved to be insightful and confirmed Pearson's commitment to improvement of their systems and procedures. That would seem to be encouraging news for the Georgia DOE, despite the occasional errors that seem plague the EOCT and GHSGT programs[3]. Certainly none of the errors were due to serious negligence on Pearson's and every attempt appears to have been made to implement changes to fix the problems. Nonetheless, we have three recommendations with commentary. These recommendations more or less reiterate themes from the Phase 1 and 2 PPSA reports.

4.1 Focus on Proactive, Systemic Changes.

Although the root cause analyses (RCAs) discussed in prior PPSA reports provide remedial, *reactive* steps to find and correct problems, the determination of solutions tend to be idiosyncratic fixes, rather than broad-based, proactive and systemic changes. The root cause analyses amount to identifying a hole in the bottom of a sinking sail boat when the overall problem may be design flaw in the rudder and hull, combined with a navigator that cannot seem to avoid rocky shallows. Plugging the hole fixes the immediate problem, but only systemic change will prevent it or other problems from happening again. The same can be said for the RCAs. They are sufficient fix a particular problem by careful retrospection, but do not include a step back to proactively re-examine the entire system. Some of the automated systems and processing changes outlined in Pearson's response to the Phase 2 EPR report suggest a move toward a more proactive response, which we whole-heartedly endorse.

4.2 Implement Tangible Reconciliation Reports and Other QC Documentation

Pearson still appears to not have fully implemented a systematic way of generating reconciliations and QC outputs for each phase of processing. This includes verifying inputs as well as outputs. The simple proof of this statement rests with the fact that such outputs should be readily available, on-demand (e.g., as requested for the EPR phase of the PPSA). For example, there is no evidence of there being a master file with prescribed naming conventions controlling inputs and outputs for the item analyses, calibrations, and scoring. All interim analysis files should be included and all analysis coding should be auto-generated to prevent analysts from inadvertently

[3] On-going personal communications between CART and Georgia DOE staff have indicated other minor problems that occurred as late as May 2010.

inputting/outputting the wrong files or worse, overwriting data files by running an old query—something that did actually happen in Summer 2009.

In addition, reconciliations should be generated for every data file and checked for accuracy. Although many of the analyses undertaken by Pearson are verified independently by two psychometricians, there was no evidence presented by Pearson that the data files are guaranteed to be correct from every possible viewing angle. Luecht (2010)[4] described the capabilities of a comprehensive QC system for foreign language tests as follows.

"Designing easy-to-use data manipulation queries and table extracts/report functions that conveniently generate different views of the data for specific analysis purposes is crucial. Too often, excellent psychometric and statistical analysts are frustrated at not being able to get the data in a way (view) that serves their purpose. Although not specifically discussed in this paper, data manipulation software was written to work off original, de-normalized source files. The data manipulation software not only prepared the three data views [noted earlier], but accounted exactly for all partial and corrupted records, precisely managed duplicates by matching on examinee identifiers, rater identifiers, and response strings, and flagged any other anomalies in the data stream. The software further automatically generated all analysis scripts/control files for the item analyses, inter-rater agreement analyses, and test-score analyses." (p. 17).

The point is that it is entirely feasible to create a scalable, automated or semi-automated examination processing system to avoid "operator error" as much as possible. This does not mean taking technical analyses out of the hands of professional psychometricians. To the contrary, it means building a fool-proof system of generating and checking the data that the psychometricians use, and ensuring that their outputs are faithfully stored in appropriate single-source data repositories. There should be automated checks on counts and other descriptive results for every output file or report generated for particular purposes. Furthermore, no report, item statistics, or scores file should be released until checked, signed off, and locked down. Summaries of aberrant conditions, data anomalies, etc. should produce active locks on out-going data streams, reports, or other information until every issue is resolved. This is a *proactive* design at its finest. While Pearson has provided some evidence that it is working on system changes that might provide many of these automated reconciliation and QC

[4] Luecht, R. M. (2010, Apil). *Some Small Sample Statistical Quality Control Procedures for Constructed Response Scoring in Language Testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, Colorado.

capabilities, until that system is in place and evidence provided that it is working, our recommendation remains "under consideration" versus "implemented".

4.3 Get Serious About Software Version Controls and Lock-Downs

Regardless of how talented a software programmer or analyst is, the reality is that most errors tend occur due to human intervention than to failures of software. Granted, a software programming error may make the same mistake over and over again. But, once fixed and recompiled, the problem should be gone—unless, of course, the compiled version does not fully replace the problem code in every instance or a user is given the latitude to make changes to certain set-ups or parameters. Automated code generators with master file naming conventions implemented for auxiliary software such as item analysis, calibration, or scoring software can help. But, in addition, a systemic top-down view of system redesign and modification should mandate formal version controls, with requisite lock downs of every change and formal sign-offs should be required, no matter how small a change is implemented.

Past errors and problems with respect to the Georgia EOCT and GHSGT underwent the root cause analyses that almost invariably discovered human errors. As alluded to in the DMR and EPE reports (Phases 1 and 2), every process that involves human interactions—especially those of software programmers and analysts—should require independent verification.

When there are immediate issues or problems requiring on-the-fly fixes, the QC group also needs to *empirically* demonstrate and document how nonstandard fixes are handled outside of the regular system (versus changed within the operational system) and then only implemented and locked down when fully justified, verified via test cases, and entered in the version-control system. A top-down review of the changes should also be mandatory to evaluate whether or not a more efficient change elsewhere in the system could automate the procedures or QC mechanisms involving the target human or software process or outputs.

Pearson's responses to the Phase 1 and 2 audit reports suggested that the company is making important strides in this direction. There term "automated" was used to describe many of their next-generation systems components. However, those systems are not in widespread use now. Indeed, the company's Asset Management System (AMS) may be the primary system intended to monitor software and processing

system changes. However, the AMS needs to be more than merely a passive reporting and documentation system. It needs to *enforce* version controls and actively restrict critical operations until all flags or other aberrant results are resolved and signed off.

4.4 Formalize Additional, Actionable QC Metrics and Readily Accessible Reports.

The Georgia Quality Action Plan (GQAP) identifies some rather high-level metrics that have been put into place or that are planned for future implementation. The metrics are "high-level" insofar as not having apparent counts or other actionable quantities that would signal improvement or at least maintaining a benchmark over time. In that regard, more specificity would be useful in defining those metrics. Furthermore, an on-demand QC tracking reports should made available to the Georgia DOE, outlining progress on each of the metrics shown in Appendix B.

The summary QC tracking and reporting metrics could also include some of the following: (a) a reconciliation summary of test forms and test booklets and unaccounted-for booklets (assumed destroyed); (b) a reconciliation summary of answer sheets sent, received, and unaccounted for (assumed destroyed); (c) a summary of scanning integrity, scanning errors/corrections and non-duplicate data records encoded post scanning; (d) a summary of answer key changes and re-scoring impact; (e) reconciliation of scores generated and posted to Georgia's servers (possibly with file check-sum information); and (f) summaries of aberrant and nonstandard processing conditions. Note that, if the values for these types of metrics are largely automated, the human resource time typically associated with auxiliary data collection (e.g., manually filling in production reports or spreadsheets) can be minimized.

Appendix A: RTPIA Visit Agenda (Prepared by Pearson)

AGENDA

CART Audit

May 17, 2010
9:00 a.m. – 2:30 p.m.

9:00 a.m. – 9:30 a.m.	**Introduction** Program Team and Organizational Quality- Shannon Still, Michelle Klingeman, Chris Skapyak	Shannon Still Michelle Klingeman Chris Skapyak
9:30 a.m. – 10:30 a.m.	**Quality Plan Review**	Chris Skapyak Dadi Narasimha
10:30 a.m. – noon	**"Live Processing" Review**	Chris Skapyak Michelle Klingeman
Noon – 1:00 p.m.	**Software Development**	Brad Russell Dadi Narasimha
1:00 p.m. – 1:30 p.m.	**Program Management**	Shannon Still Michelle Klingeman
1:30 p.m. - 2:30 p.m.	**CART Requests**	Chris Skapyak Shannon Still
2:30 p.m.	**Closing**	All

Appendix B: Pearson Response to the CART Phase 1 DMR Report and Pearson's Georgia Quality Action Plan (GQAP)

Pearson Response to the Center for Assessment and Research Technology Findings

Pearson appreciates the opportunity to respond to the detailed audit findings completed by the Center for Assessment and Research Technology (CART) – November 9, 2009. As an organization, Pearson is committed to the continuous improvement of our quality, processes, and systems. Responding to the audit has been an insightful and constructive process, instrumental in strengthening our resolve to maintain the integrity, reliability, and accountability of all of our organizational systems and processes.

Though we believe our processes and systems operate at or above industry standard, we found a great deal of value in the recommendations and identified considerable opportunity to enhance our level of service to Georgia and other state clients. The report recommendations highlight valid and important areas for improvement; some areas of improvement Pearson initiated prior to the audit as part of our ongoing commitment to quality and some improvements are a direct result of the audit. Listed below are the audit recommendations followed by Pearson's responses.

In addition to our responses to the CART recommendations below, we have included a detailed and comprehensive Georgia Quality Action Plan that specifies action, ownership, and timeline for implementation of our quality initiatives and procedures. We strongly believe that this detailed plan will be an important tool to drive, monitor, and evaluate quality for the Georgia assessment programs for both Pearson and the Georgia Department of Education (GaDOE).

We will look forward to a formal audit debrief to further discuss our response and action plan with the GaDOE and CART. We believe such a meeting will allow for meaningful and insightful dialogue– ultimately resulting in a clear understanding of the quality plan and initiatives implemented to deliver Georgia's assessment programs.

CART Recommendation #1 - *Our first recommendation is that Pearson develops a larger array of automated mechanisms—that is, software applications and routines—for reconciling every aspect of the data throughout the processing cycles.*

Pearson Response: Pearson concurs that automated mechanisms for software applications and routines often increase both quality and productivity. While Pearson follows the industry standard of design, development, unit testing, system testing, end to end testing, and an independent production validation step through our Assessment and Information Quality

(AIQ)group we continue to look for ways to automate our work. Pearson can point to a number of places where we have implemented automation to improve both quality and customer response time. Some examples include:

- CAWA (CA Workflow Automation) system: facilitates the flow from one job/one system to another. This will be implemented on the Georgia program for the 2010 summer administration resulting in a reduction in the number of manual interventions.
- Messaging between systems: communicates data information for the purposes of reconciliation between systems. These activities ensure that what was sent by one system was received by the next system. This capability is contained within the CAWA functionality which will be implemented in summer 2010 for Georgia.
- Implementation of cross checking opportunities, where applicable. We are evaluating all opportunities to utilize this functionality on the Georgia program. Examples of some of these opportunities could include comparing ISR data to roster data, comparing summary role up data to the actual summary data.
- Ncount Comparisons - verify that all of the enrollment quantities within a specified packaging and distribution file are correct, based upon the enrollment values found within the Enrollment Master file. This is currently taking place for Georgia programs now.
- F500 Val: validates the F500 file against the project specification form and has already been implemented.
- Summary/Aggregation tool: a java based application that helps automate the Summary File creation and validation process. This will be implemented for future administrations and discussed with GADOE during the audit debrief.

Additionally, Pearson has two separate automation groups that are constantly looking at ways of utilizing automation to improve our processes.

The first of these two groups is the AIQ automation group. The AIQ Automation group has defined coding standards and is in the process of creating and/or maintaining automation tools that support Pearson assessment programs. When the group was formed in the 2nd half of 2009, it developed coding standards, naming conventions, and standard automation tools for use in test administrations. To date, the group has developed over forty-three (43) automation applications. Some of the applications are administration or project-specific and others are global. Some of the global tools include creating test data for quality testing, validating our ePen files (files of student constructed responses for image based scoring), and validating F500 files.

The second automation group, Software &Technology (S&T) Automation supports automation testing processes and standards across S&T's product lines such as Pearson Access ─ and provide a range of automation services including:

- ***Functional Automation*** – Automated activities performed on one system to verify and validate the specified functionality is performing as expected.
- ***Integration Automation*** – Automated activities performed to verify and validate that the integration points between two or more applications retain the integrity of the data being passed from one system to the next.
- ***Regression Automation*** – Automation to verify that existing system functionality is not ill-affected when new or modified code is introduced by development.
- ***Custom Tool Development*** – Design, develop and implement tools to drive quality and improve efficiencies.

Discrepancies should not be passively reported and then possibly ignored[5]*. Rather, all discrepancies—depending on their severity as determined by predetermined policies—should always generate automated flags that: (a) stop further production or at least activate a series of prescribed steps (e.g., activate an exception handling agent); (b) require resolution with active sign-off; and (c) mandate secondary follow-up checking to ensure that the resolution and sign-offs were implemented."*

Pearson Response: Pearson has established policies for "discrepancies" that could potentially occur in the different stages of project delivery. Examples include:

1. Flags during production processing; project "flags" are specified and documented in configuration documents. These specifications determine automated checks which flag records. When an item is "flagged" due to a data value, the item is reviewed to determine whether there is a mis-key, a format or structural problem with the item, a content issue such as an ambiguous question, or whether the item appears fine. Measured Progress (MP) or the GADOE Content Specialists may be consulted to help determine the resolution of the issue.
2. Automated editing during imaging: our image editing system automatically alerts specific cases and will stop processing for that batch until further analysis can be performed. These edits are all based on defined specifications. Once resolution is

[5] Many of Pearson's exception logs and reports appear to be entered into Microsoft Word or Excel documents. Although that is fine for human review, it is not clear that these files are linked to requisite actions or that automatically halt further processing until the discrepancies are resolved. For example, an item key validation discrepancy flag should requires concrete resolution and action to remove the flag—not just passive human review of an Excel file and [possible] sign-off—before the item analysis or scoring can proceed.

reached for the flagged records or the alerted records, the batches or records can be re-introduced for further processing.

3. Quality Testing: Discrepancies found during testing are handled differently. As a discrepancy is found in the code in relation to the requirements document, a team track is entered specifying the discrepancy. This is sent to the IT Project Manager (ITPM) initially and then passed on to the appropriate individuals for resolution. Once resolution is reached, the changes are made (either in the documentation or the code); the changes are tested by testing groups and signed off. This process requires multiple documented sign-offs.

CART Recommendation #2 - *Our second recommendation is that Pearson make every attempt to implement procedures that require changes to be directly performed in the master database sources (e.g., changing answer keys, item images, etc.), rather than allowing modifications to extracts and or "detached" versions of the data records."*

Pearson Response: In our current publishing approach, the Pearson Item Tracker is a blend of a development tool and reference repository. At the time we developed the tools, the technology was not available to drive publishing from a single repository. We consider the current "database of record" to be the fileserver and files maintained by the Publishing Operations group since they have the actual items and forms published.

Pearson concurs with the recommendation that an appropriately controlled single item repository ("master database") such as an XML-driven solution is a best practice that the industry is moving toward. Pearson identified this as part of our strategic vision about one year ago and has begun the design approach for our future Asset Management System (AMS). We are currently capturing requirements for the system design and have a development schedule outlined. A more detailed description of the Asset Management System is included below for your review.

Until the AMS system is completely developed, Pearson will continue to follow established, industry standard procedures as communicated during the CART audit. We will also incorporate the additional quality checks detailed in the quality plan (Addendum A).

CART Recommendation #3 - *Our third recommendation is to move to database representations of as much of the processing as possible, as opposed to using scripts, SQL code with hard-coded file or table names, etc... For example, scripts or program code should use tokens for file names and other set-ups. "*

Pearson Response: Pearson agrees with this recommendation, and believes it is in line with the direction the industry is moving. Our technology development strategy includes design standards that move our technology platforms towards more parameterized and configurable solutions. Many of these elements are actively being incorporated into the AMS tool.

Any changes to the answer keys are managed under the Pearson change management process. Following this process forces the lookup tables, the OSA configuration file and script to be kept in sync. The change must be identified to IT via a formalized change control process and is tracked to closure.

We will further review opportunities to look at hard-coded files and table names as we make improvements to this process.

CART Recommendation # 4 - *Our final recommendation recognizes that it is virtually impossible to fully automate any system. That being said, where ever humans touch the systems or data (e.g., manually preparing scripts, running extract queries/SQL, or entering field names in analysis records in a database),implement mandatory 2Is with active sign-offs, as noted in Section 3.5.6. Also, all results, from equating to score reporting should employ the reconcile-to-expectation (R2E) principal, again, with active sign-off. "*

Pearson Response: Pearson recognizes the importance of IT automation as well as understanding the intricacies of a project which automation may not be able to "check". We agree with this recommendation and as part of the standard software development process, where humans touch systems or data Pearson implements the following peer review processes:

- *Software Code;* Software code is reviewed by multiple peers. Materials are distributed in advance, updates are identified and tracked to closure through the formal peer review process.
- *Test Plans:* Test Plans are reviewed by multiple peers. This peer review process has now been formalized and is taking on the Georgia program. Materials are distributed in advance, updates are identified and tracked to closure through the formal peer review process
- *Readiness Reviews:* Readiness reviews will be completed prior to handoff to each stage of testing as well as prior to production.
- *Final Review* – A final review of the production data files and reports is completed prior to customer hand-off.

From a psychometric perspective, Pearson psychometricians run SAS programs on flat file inputs (files provided by IT) and apply the 2I's principle. We save the input files, the SAS code used to create the IDMs, the log files from the SAS runs, and the output files from the SAS runs. We use versioning control for inputs, outputs and processing code. Thus, we believe our procedures are consistent with this recommendation although our process does not include the dynamic data system referred to in the recommendation.

Overview of the Asset Management System (AMS)

Pearson is well into the process of implementing our Asset Management System (AMS). This will be a phased implementation over the next several months. The AMS system will address many of the issues highlighted in the audit findings. Project AMS is the assessment development workflow improvement tool designed to reform the way in which we design, develop, construct and deploy educational assessments. Project AMS aims to construct a consistent, lean, cost-effective, quality focused assessment development workflow. Pearson AMS impacts the work currently being performed in Austin, Iowa City, Owatonna and San Antonio within the shared service functions of Test Measurement Research Services (TMRS), Publishing Operations, and IT.

With this system we will introduce a level of automation and built in quality control mechanisms that specifically address the majority of the opportunities presented in the audit findings. AMS will rely on a “gold standard” master for all assets. An asset is defined as an item, art, and/or the combination of one of these and the associated Meta data. These items will be created as .xml objects that make them available to use in either paper based or electronic testing. Each item will have a unique identification number with absolutely no opportunity for replication. The test map will be utilized to compose each administration specific grouping of assets. The requirements for test construction tools will be defined, including the functionality of the tool that will be embedded in the content management system, as well as the requirements of the system to support tools that need to be developed outside of the system. This workgroup will also standardize the item-level and test-level psychometric data that will be produced and imported into the system.

Further Commitment to Quality

As an additional commitment to quality, Pearson is including an overall Quality Action Plan (Addendum A). This Quality Action Plan details elements of our quality plan that are already in place as well as those added specifically for the Georgia program.

Addendum A